

VISIT...

LANZAROTE
Caliente.COM

KIRK-OTHMER

**ENCYCLOPEDIA OF
CHEMICAL
TECHNOLOGY**

EDITING - by **WATCHER**

SUPPLEMENT

FOURTH EDITION

**AEROGELS TO
XYLYLENE POLYMERS**

CONTENT INDEX (vol 26, Supplement)

(with hyperlinks)

Edited by <http://www.watcherworld.narod.ru/>

AEROGELS	1
AQUACULTURE	12
BIOREMEDIATION	24
BIOSEPARATIONS	42
CHIRAL SEPARATIONS	59
FURAN DERIVATIVES	74
HIGH PURITY GASES	87
HORMONES, ADRENAL-CORTICAL	93
LITHOGRAPHIC RESISTS	113
MICROBIAL AND VIRAL FILTRATION	139
MICROEMULSIONS	147
MOLECULAR MODELING	157
MOLECULAR RECOGNITION	174
NANOTECHNOLOGY	199
NITRILES	217
RECYCLING, PLASTICS	229
SEMICONDUCTORS, ORGANIC	236
SMART MATERIALS	249
SOLVENT RECOVERY, CONDENSATION	254
SONOCHEMISTRY	255
SURFACE AND INTERFACE ANALYSIS	268
THERMOGRAPHY	290
UREA	297
WATER-SOLUBLE POLYMERS	312
WEIGHING AND PROPORTIONING	324
WELDING	341
WHEAT AND OTHER CEREAL GRAINS	351
WINE	363
WOOD-BASED COMPOSITES AND LAMINATES	377
XANTHENE DYES	398
XYLENES AND ETHYLBENZENE	410
XYLYLENE POLYMERS	427

AEROGELS

Aerogels, solid materials that are so porous that they contain mostly air, were first prepared in 1931. Kistler used a technique known as supercritical drying to remove the solvent from a gel (a solid network that encapsulates its solvent) such that, as stated in his paper in *Nature*, "no evaporation of liquid can occur and consequently no contraction of the gel can be brought about by capillary forces at its surface (1)." In the six decades since then, there have been significant advances made both in the use of new precursors to form gels and in the removal of solvent from them. These advances have greatly simplified the preparation of aerogels and, in turn, improved their economic viability for commercial applications. Increasingly an aerogel is defined in terms of its properties and not the way in which it is prepared.

Almost all applications of aerogels are based on the unique properties associated with a highly porous network. Envision an aerogel as a sponge consisting of many interconnecting particles. The particles are so small and so loosely connected that the void space in the sponge, the pores, can make up for over 90% of the sponge volume. As an example, a silica aerogel contains particles that are of the order of 10 nm and each particle is connected to two or three other particles on average. Such a material has a typical density of about 100 kg/m³ and accessible surface area of about 1000 m²/g.

The ability to prepare materials of such low density, and perhaps more importantly, to vary the density in a controlled manner, is indeed what make aerogels attractive in many applications such as thermal insulation, detection of high energy particles, and catalysis. Thus, this article begins with a discussion of sol–gel chemistry used to form the wet gel. The intention is not to discuss in detail the fundamental chemistry involved, but to provide the basic principles that explain the formation of a porous network and the effect of preparative variables on its microstructure (often referred to as nanostructure because the relevant length scale is on the order of nanometer), ie, factors that impact on density. The general applicability of these principles are illustrated by examples of inorganic, organic, and inorganic–organic gels.

The section on preparation and manufacturing continues to focus on the microstructure of a gel, in particular its evolution during the removal of solvent. In addition to the original supercritical drying used by Kistler, there are now safer and more cost-effective methods that do not involve supercritical drying. More importantly, these methods ensure that the products possess the defining characteristics of an aerogel (ultrafine pores, small interconnected particles, and high porosity).

This article also aims at establishing the structure–property–application relationships of aerogels. Selected examples are given to show what some desirable properties are and how they can be delivered by design based on an understanding of the preparation and preservation of a gel's microstructure.

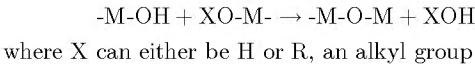
Sol–Gel Chemistry

Inorganic Materials. Sol–gel chemistry involves first the formation of a sol, which is a suspension of solid particles in a liquid, then of a gel, which is a diphasic material with a solid encapsulating a solvent. A detailed description of the fundamental chemistry is available in the literature (2–4). The chemistry involving the most commonly used precursors, the alkoxides (M(OR)_m), can be described in terms of two classes of reactions:

Hydrolysis



Condensation



The important feature is that a three-dimensional gel network comes from the condensation of partially hydrolyzed species. Thus, the microstructure of a gel is governed by the rate of particle (cluster) growth and their extent of crosslinking or, more specifically, by the *relative* rates of hydrolysis and condensation (3).

The gelation of silica from tetraethylorthosilicate (TEOS) serves as an example of the above principle. Under acidic conditions, hydrolysis occurs at a faster rate than condensation and the resulting gel is weakly branched. Under basic conditions, the reverse is true and the resulting gel is highly branched and contains colloidal aggregates (5,6). Furthermore, acid-catalyzed gels contain higher concentrations of adsorbed water, silanol groups, and unreacted alkoxy groups than base-catalyzed ones (7). These differences in microstructure and surface functionality, shown schematically in Figure 1, result in different responses to heat treatment. Acid- and base-catalyzed gels yield micro- (pore width less than 2 nm) and meso-porous (2–50 nm) materials, respectively, upon heating (8). Clearly an acid-catalyzed gel which is weakly branched and contains surface functionalities that promote further condensation collapses to give micropores. This example highlights a crucial point: *the initial microstructure and surface functionality of a gel dictates the properties of the heat-treated product.*

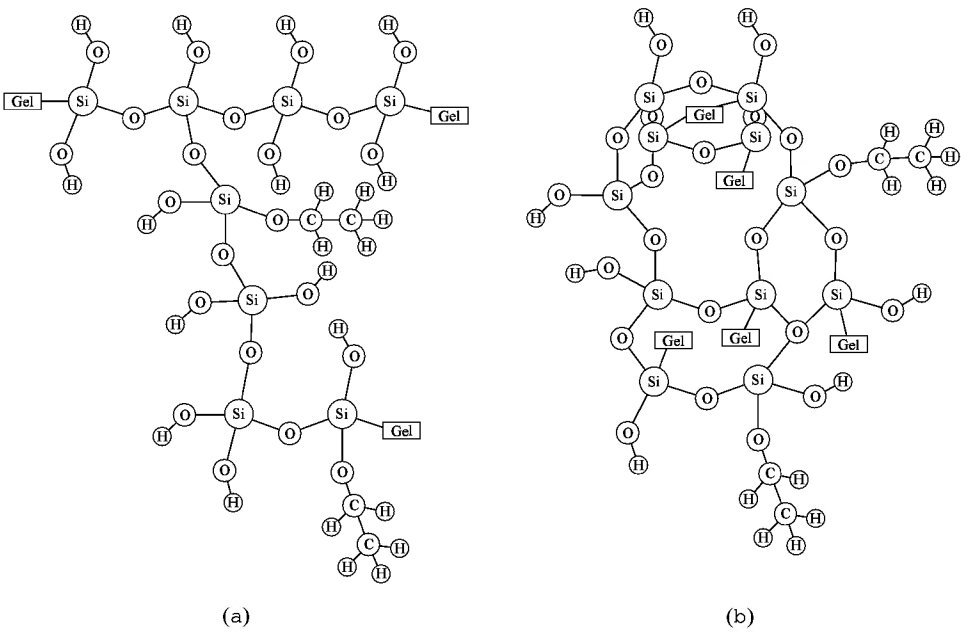


Fig. 1. Schematics of (a) acid-catalyzed and (b) base-catalyzed silica gels showing the differences in microstructure and surface functional groups. Reproduced from Ref. 7.

Courtesy of the American Ceramic Society.

Besides pH, other preparative variables that can affect the microstructure of a gel, and consequently, the properties of the dried and heat-treated product include water content, solvent, precursor type and concentration, and temperature (9). Of these, water content has been studied most extensively because of its large effect on gelation and its relative ease of use as a preparative variable. In general, too little water (less than one mole per mole of metal alkoxide) prevents gelation and too much (more than the stoichiometric amount) leads to precipitation (3,9). Other than the amount of water used, the rate at which it is added offers another level of control over gel characteristics.

The principles discussed so far are valid for silicates as well as nonsilicates, although there are more data available for the former. The alkoxides of transition metals tend to be more reactive than silicon alkoxides because both hydrolysis and condensation are nucleophilic substitution reactions and their cations have a more positive partial charge than silicon (3). This difference in reactivity presents both a challenge and an opportunity in the preparation of two-component systems. In a two-component system, the minor component can either be a network modifier or a network former. In the latter case, the distribution of the two components, or mixing, at a molecular level is governed by the *relative* precursor reactivity. Qualitatively good mixing is achieved when two precursors have similar reactivities. When two precursors have dissimilar reactivities, the sol-gel technique offers several strategies to prepare well-mixed two-component gels. Two such strategies are prehydrolysis (10), which involves prereacting a less reactive precursor with water to give it a head start, and chemical modification (11), which involves slowing down a more reactive precursor by substituting some of its alkoxy groups with bulkier, less reactive groups such as acetate. The ability to control microstructure *and* component mixing is what sets sol-gel apart from other methods in preparing multicomponent solids.

Organic Materials. The first organic aerogel was prepared by the aqueous polycondensation of resorcinol with formaldehyde using sodium carbonate as a base catalyst (12). Figure 2a shows that the chemistry is similar to the sol-gel chemistry of inorganic materials. Subsequent to the reaction between resorcinol and formaldehyde, the functionalized resorcinol rings (ie, those possessing $\text{-CH}_2\text{OH}$ groups) condense with each other to form nanometer-sized clusters, which then crosslink to form a gel. The process is influenced by typical sol-gel parameters such as pH, reactant concentration, and temperature. The resorcinol-catalyst molar ratio turns out to be the dominant factor that affects the microstructure of the gel (13).

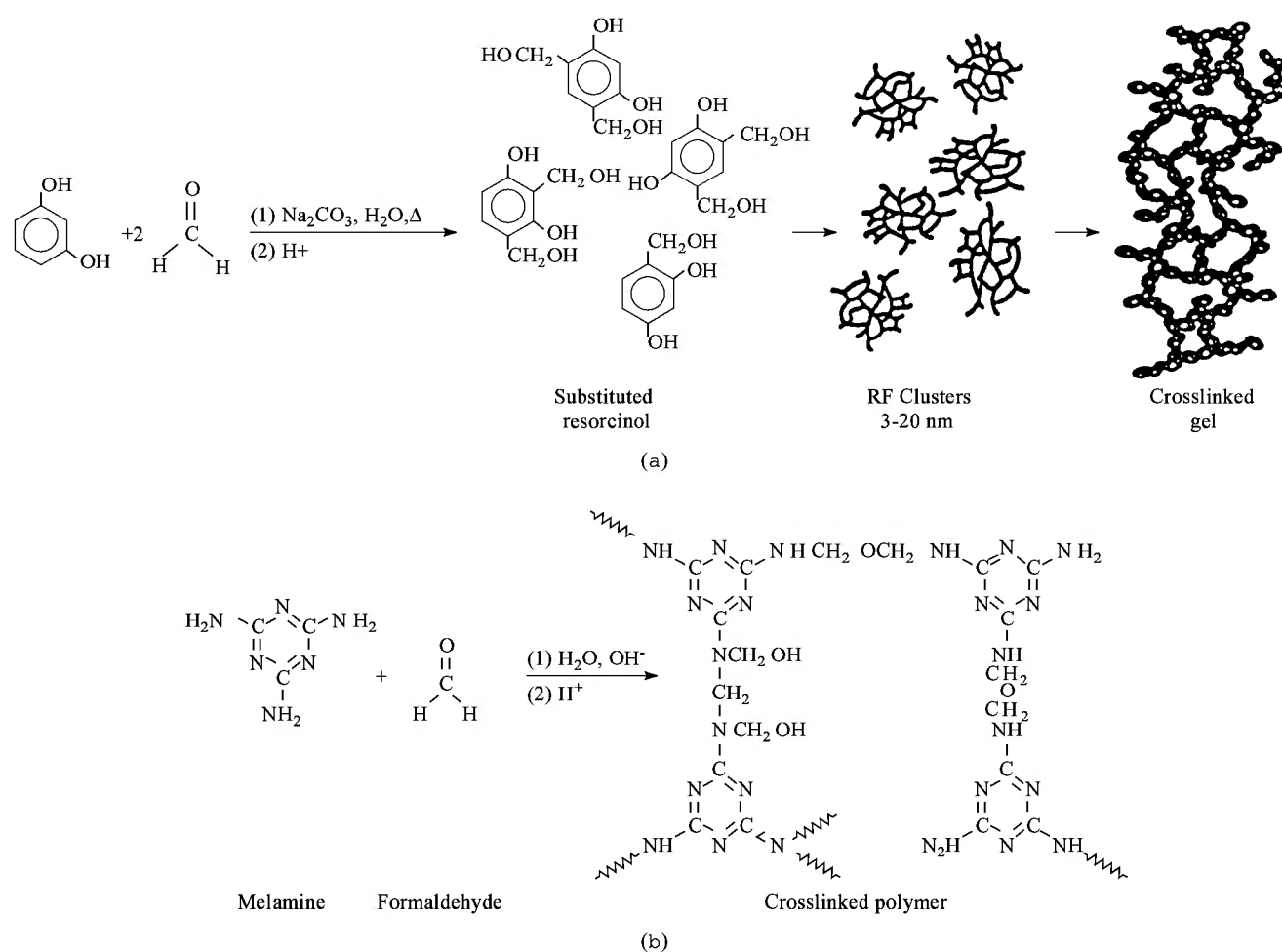


Fig. 2. The sol-gel polymerization of resorcinol with formaldehyde (a) and (b) of melamine with formaldehyde. Reproduced from Refs. 13 and 14, respectively.

Courtesy of the Materials Research Society.

Resorcinol-formaldehyde gels are dark red in color and do not transmit light. A colorless and transparent organic gel can be prepared by reacting melamine with formaldehyde using sodium hydroxide as a base catalyst (14). The chemistry, shown in Figure 2b, involves first the formation of $\text{-CH}_2\text{OH}$ (hydroxymethyl) groups and then the formation of $\text{-NHCH}_2\text{NH-}$ (diamino methylene) and $\text{-NHCH}_2\text{OHC}_2\text{H-}$ (diamino methylene ether) bridges through cross-linking reactions. The solution pH is again a key parameter, it is necessary to add hydrochloric acid in the second part of the process to promote condensation and gel formation.

Both the preparations of resorcinol-formaldehyde and melamine-formaldehyde gels are aqueous-based. Since water is deleterious to a gel's structure at high temperatures and immiscible with carbon dioxide (a commonly used supercritical drying agent), these gels cannot be supercritically dried without a tedious solvent-exchange step. In order to circumvent this problem, an alternative synthetic route of organic gels that is based upon a phenolic-furfural reaction using an acid catalyst has been developed (15). The solvent-exchange step is eliminated by using alcohol as a solvent. The phenolic-furfural gels are dark brown in color.

Carbon aerogels can be prepared from the organic gels mentioned above by supercritical drying with carbon dioxide and a subsequent heat-treating step in an inert atmosphere. For example, the following heating schedule has been used in flowing nitrogen (13,15): 22 to 250°C in 2 h, held at 250°C for 4 h, 250 to 1050°C in 9.5 h, held at 1050°C for 4 h, and cooled to room temperature in 16 h. This treatment results in a volume shrinkage of the sample of 65–75% and a mass of 45–50%. Despite these changes, the carbon aerogels are similar in morphology to their organic precursors, underscoring again the

importance of structural control in the gelation step. Furthermore, changing the sol–gel conditions can lead to aerogels that have a wide range of physical properties. As a specific example, surface area for the phenolic–furfural aerogels is about $385 \pm 16 \text{ m}^2/\text{g}$ over a density range of $100\text{--}250 \text{ kg}/\text{m}^3$, whereas the corresponding carbon aerogels have surface areas of $512 \pm 40 \text{ m}^2/\text{g}$ over a density range of $300\text{--}450 \text{ kg}/\text{m}^3$.

Inorganic–Organic Hybrids. One of the fastest growing areas in sol–gel processing is the preparation of materials containing both inorganic and organic components. The reason is that many applications demand special properties that pure materials can seldom provide. The combination of inorganic and organic materials is, thus, an attractive way to deliver materials that have desirable physical, chemical, and structural characteristics. In this regard, sol–gel chemistry offers a real advantage because its mild preparation conditions do not degrade organic polymers, as would the high temperatures that are associated with conventional ceramic processing techniques. The voluminous literature on the sol–gel preparation of inorganic–organic hybrids can be found in several recent reviews (16–20) and the references therein; only a qualitative sketch is given in this section.

There are several ways to classify inorganic–organic materials, all of which depend on the strength of interaction between the two components. The interaction can range from nonexistent or weak, such as organic species embedded (or entrapped) in an inorganic network, to very strong, such as systems involving covalent bonds. For weakly interacting systems, the general observation is that most molecules can be entrapped in sol–gel matrices and, once entrapped, these molecules retain most of their characteristic physical and chemical properties (18). The entrapment of bioorganic materials such as enzymes, whole cells, antibodies, and other proteins (21) falls into this category.

For strongly interacting systems, chemical bonding can be induced by using functionalized precursors. There are three basic types of precursors: inorganically functionalized preformed organic polymers, organically functionalized oxides, and precursors containing both inorganic and organic functional groups. Examples of commonly used precursors are organofunctional metal alkoxides $(\text{RO})_n\text{--E--X--A}$ (19) and bridged polysilsesquioxanes (20). In $(\text{RO})_n\text{--E--X--A}$, A is a functional organic group and X is a hydrolytically stable spacer linking A and the metal alkoxide which provides the inorganic function. The use of such precursors has allowed the control of very small domain sizes, often in the nanometer range, in the preparation of inorganic–organic materials (16,19,20). The challenge is to achieve a high degree of mixing of the two phases, thus, enabling the manipulation of interfacial properties at a molecular level. Another promising strategy to provide better homogeneity between the two phases is to form the inorganic and organic phases simultaneously, leading to what is known as a simultaneous interpenetrating network (16).

All the methods described so far involve introducing the organic phase prior to the formation of a solid phase. There is an interesting alternative to prepare a nanocomposite by a sequential approach (22). In this approach a silica aerogel was first prepared, then carbon was deposited in it by the decomposition by hydrocarbons (eg, methane, acetylene) at a temperature range of $500\text{--}850^\circ\text{C}$. This example demonstrates the feasibility of preparing hybrid materials by depositing an inorganic or organic phase onto an organic or inorganic substrate, respectively.

Preparation and Manufacturing

Supercritical Drying. The development of aerogel technology from the original work of Kistler to about late 1980s has been reviewed (23). Over this period, supercritical drying was the dominant method in preparing aerogels and, for this reason, aerogels are synonymous with supercritically dried materials. As noted earlier, supercritical drying could be an insufficient definition of an aerogel because it might not lead to the defining characteristics. Kistler used inorganic salts and a large amount of water in his work (1,24), making the subsequent salt removal and solvent-exchange steps time-consuming (water has to be removed because it would dissolve the gel structure at high temperatures). A significant time savings came when alkoxides were used as precursors in organic solvents, thereby requiring a minimum amount of water and eliminating the tedious solvent-exchange step (25,26). The introduction of carbon dioxide, which has a lower critical temperature and pressure than alcohols, as a supercritical drying agent allowed the drying step to be done under milder conditions and improved its safety (27). The development of a semicontinuous drying process further facilitated the preparation of aerogels (28). Together these advances, summarized in Table 1, have made possible the relatively safe supercritical drying of aerogels in a matter of hours. In recent years, the challenge has been to produce aerogel-like materials without using supercritical drying at all in an attempt to deliver economically competitive products. This topic will be discussed in more detail later.

Table 1. Important Developments in the Preparation of Aerogels

Decade	Developments
1930	Using inorganic salts as precursors, alcohol as the supercritical drying agent, and a batch process; a solvent-exchange step was necessary to remove water from the gel.
1960	Using alkoxides as precursors, alcohol as the supercritical drying agent, and a batch process; the solvent exchange step was eliminated.
1980	Using alkoxides as precursor, carbon dioxide as the drying agent, and a semicontinuous process; the drying procedure became safer and faster. Introduction of organic aerogels.
1990	Producing aerogel-like materials without supercritical drying at all; preparation of inorganic–organic hybrid materials.

As stated earlier, the main idea behind supercritical drying is to eliminate the liquid–vapor interface inside a pore, thereby removing the accompanying capillary pressure which acts to collapse a gel network. The value of this approach is demonstrated by the fact that aerogels do have higher porosities, higher specific surfaces areas, and lower apparent densities than xerogels, materials that are prepared by evaporative drying. However, it is incorrect to think that a gel remains static during supercritical drying. Rather, supercritical drying should be considered as part of the aging process, during which events such as condensation, dissolution, and reprecipitation can occur. The extent to which a gel undergoes aging during supercritical drying depends on the structure of the initial gel network. For example, it has been shown that a higher drying temperature changes the particle structure of base-catalyzed silica aerogels but not that of acid-catalyzed ones (29). It is also known that gels that have uniform-sized pores can withstand the capillary forces during drying better because of a more uniform stress distribution. Such gels can be prepared by a careful manipulation of sol–gel parameters such as pH and solvent or by the use of so-called drying control chemical additives (DCCA) (30). Clearly, an understanding of the interrelationship between preparative and drying parameters is important in controlling the properties of aerogels.

The most widely studied supercritical drying variable is temperature simply because different solvents have different supercritical temperatures. Specifically, since alcohols have higher supercritical temperatures than carbon dioxide, there have been many recent reports on the effect of drying agent on the textural properties and crystallization behavior of aerogels (31–33). These results demonstrate nicely the accelerated aging that a gel undergoes at high temperatures and pressures. For this reason carbon dioxide is the drying agent of choice if the goal is to stabilize kinetically constrained structure, and materials prepared by this low-temperature route are referred to by some people as *carbogels*. In general, carbogels are also different from aerogels in surface functionality, in particular hydrophilicity, which impacts on the moisture sensitivity of these materials and their subsequent transformations under heat treatment.

It is less well known, but certainly no less important, that even with carbon dioxide as a drying agent, the supercritical drying conditions can also affect the properties of a product. For example, in the preparation of titania aerogels, temperature, pressure, the use of either liquid or supercritical CO_2 , and the drying duration have all been shown to affect the surface area, pore volume, and pore size distributions of both the as-dried and calcined materials (34,35). The specific effect of using either liquid or supercritical CO_2 is shown in Figure 3 as an illustration (36).

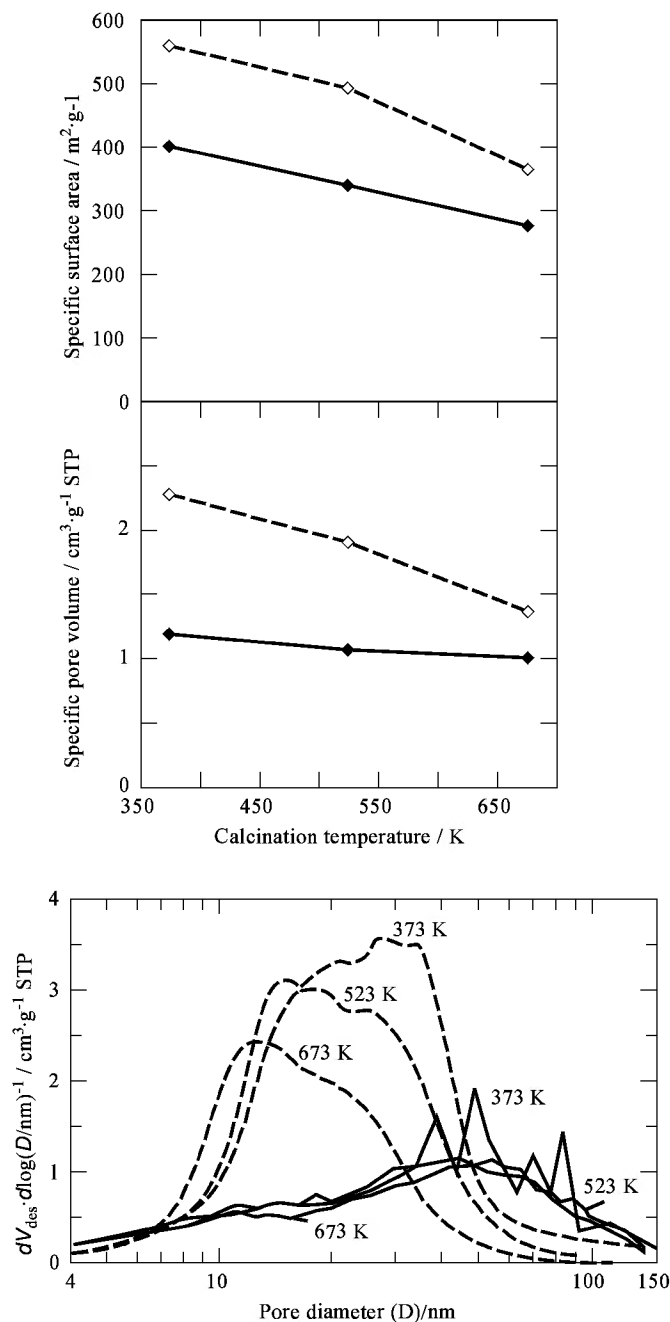


Fig. 3. Effect of using either liquid or supercritical carbon dioxide on the textural properties of titania aerogels calcined at the temperatures shown. (—), dried with liquid carbon dioxide at 6 MPa and 283 K; (---), dried with supercritical carbon dioxide at 30 MPa and 323 K. Reproduced from Ref. 36.

Courtesy of Marcel Dekker, Inc.

Other important drying variables include the path to the critical point, composition of the drying medium, and depressurization (37). The rates of heating and depressurization are especially important in the preparation of monoliths because if the pore liquid does not have sufficient time to flow out of the gel network, it could lead to excessive stresses that cause cracking (38,39). The container of a gel is by itself a source of stresses by preventing the radial expansion and flow of liquid. Quantitative models describing these phenomena are available and have been tested against experimental data for silica gels (38–40).

For some applications it is desirable to prepare aerogels as thin films that are either self-supporting or supported on another substrate. All common coating methods such as dip coating, spin coating, and spray coating can be used to prepare gel films. However, for highly porous films (ie, porosity > 75%), special care is necessary to minimize the rate of solvent evaporation both during and after gel formation. One way to do so is to perform the coating processes within an enclosure that is filled with the saturated vapor of the working solvent and a partial pressure of ammonium hydroxide that catalyzes the gelation of the films (41). The subsequent supercritical drying step can be done in either alcohol or carbon dioxide. The choice depends on the desired properties of the aerogels and, in the case of supported films, the thermal stability of the substrate materials.

In all the processes discussed above, the gelation and supercritical drying steps are done sequentially. Recently a process that involves the direct injection of the precursor into a strong mold body followed by rapid heating for gelation and supercritical drying to take place was reported (42). By eliminating the need of forming a gel first, this entire process can be done in less than three hours per cycle. Besides the saving in time, gel containment minimizes some stresses and makes it possible to produce near net-shape aerogels and precision surfaces. The optical and thermal properties of silica aerogels thus prepared are comparable to those prepared with conventional methods (42).

Ambient Preparations. Supercritical drying with alcohols incurs high capital and operating costs because the process is run at high temperatures and pressures and needs to remove a large amount of pore liquid. Carbon dioxide allows drying to be done under milder conditions but its use is limited to miscible solvents. Thus, economic and safety considerations have provided a strong motivation for the development of techniques that can produce aerogel-like materials at ambient conditions, ie, without supercritical drying. The strategy is to minimize the deleterious effect of capillary pressure which is given by:

$$P = 2\sigma\cos(\theta) / r$$

where P is capillary pressure, σ is surface tension, θ is the contact angle between liquid and solid, and r is pore radius.

The equation above suggests that one approach would be to use a pore liquid that has a low surface tension. Indeed, two-step acid–base or acid–acid catalyzed silica gels have been made, aged in ethanol or water, washed with various aprotic solvents, and finally evaporatively dried at 323 K for 48 hours and then at 383 K for 48 hours (43). The aprotic solvents used and their corresponding surface tension in N/m at room temperature (shown in

parentheses) are tetrahydrofuran (23.7×10^{-3}), acetone (23.7×10^{-3}), cyclohexane (25.3×10^{-3}), acetonitrile (29.3×10^{-3}), nitromethane (32.7×10^{-3}), and 1:4 dioxane (33.6×10^{-3}). For both the acid-catalyzed and base-catalyzed gels, it was found that increasing surface tension causes a decrease in surface area and total pore volume. However, for acid-catalyzed gels with an intermediate water wash, the micro pore volume increases with increasing surface tension. It was suggested that the water wash leads to a hydroxylated surface which, upon further condensation, gives a stiffer network and consequently a larger fraction of micropore volume. The important point is that with a pore liquid that has a sufficiently small surface tension, ambient pressure aerogels can have comparable pore volume and bulk density to those prepared with supercritical drying (Fig. 4) (44).

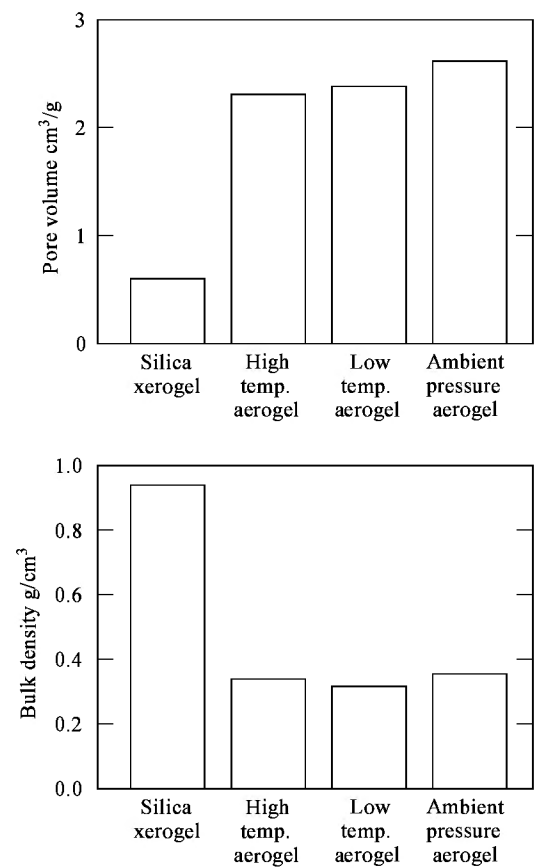


Fig. 4. Comparison of physical properties of silica xerogels and aerogels. Note the similar properties of the aerogels prepared with and without supercritical drying. Reproduced from Ref. 44.

Courtesy of the Materials Research Society.

For base-catalyzed silica gels, it has been shown that modifying the surface functionality is an effective way to minimize drying shrinkage (45,46). In particular, surface hydroxyl groups, the condensation of which leads to pore collapse, can be "capped off" via reactions with organic groups such as tetraethoxysilane and trimethylchlorosilane. This surface modification approach (also referred to as surface derivatization), initially developed for bulk specimens, has recently been applied to the preparation of thin films (47,48). The process involves the following steps: (1) prepare a base-catalyzed gel, (2) wash the gel with hexane, (3) replace the surface hydroxyls with organosilicon groups, (4) reliquify the surface-modified gel with ultrasound, (5) dip-coat the redispersed sols onto a silicon substrate, and (6) heat treat the film in air at 450°C. During drying these materials exhibit reversible shrinkage in a gradual dilation, or "spring-back," of the film. The extent of spring-back is a function of processing conditions. Films with porosity in the range of 30–99% can be prepared via a proper control of the washing, surface modification, dip-coating, and heat treatment conditions (48).

In changing surface hydroxyls into organosilicon groups, surface modification has an additional advantage of producing hydrophobic gels. This feature, namely the immiscibility of surface-modified gel with water, has led to the development of a rapid extractive drying process shown in Figure 5 (49). The basic idea is to submerge a wet gel into a pool of hot water that is above the boiling point of the pore fluid. After the pore fluid boils out of the gel, the gel floats to the surface because it is not wetted by water and can be easily recovered. Other working fluids, such as ethylene glycol and glycerol, can be used instead of water as long as they have a high boiling point and do not wet the gel. This ambient pressure process offers improved heat transfer rates and, in turn, greater energy efficiency without compromising desirable aerogel properties.

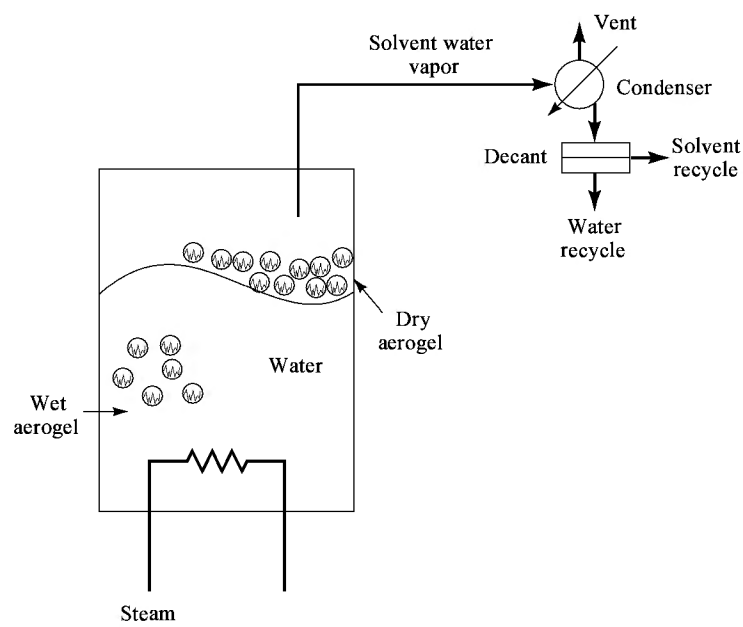


Fig. 5. Schematic diagram of an extractive drying process that produces aerogels at ambient pressure. Reproduced from Ref. 49.

Courtesy of the Materials Research Society.

Another approach to produce aerogels without supercritical drying is freeze drying, in which the liquid–vapor interface is eliminated by freezing a wet gel into a solid and then subliming the solvent to form what is known as a *cryogel*. Some potential problems are that cracks may develop in the frozen gel and sublimation, if done too fast, can melt the solvent. Cryogels of silica and nickel oxide–alumina have similar, but not identical, properties to aerogels of the same materials (50). In general, there has not been a lot of study on freeze drying and the limited data available suggest that it might not be as attractive as the above ambient approaches in producing aerogels on a commercial scale.

Properties

A detailed discussion of the properties of aerogels can be found in several recent review articles (51–55) and the references therein. This section provides a physical basis for these properties by focusing on the microstructure of an aerogel. The intent is to provide a bridge between the two previous sections, which discuss the preparative and drying parameters that affect microstructure, and the next one, which outlines the potential applications made possible by unique structural features. The emphasis is on silica aerogels because they have been the most extensively characterized.

All properties of aerogels are related to the extremely high porosity of these materials. Consider an aerogel as consisting of strings of pearls; each pearl is a structural unit and connected to only two to three other units on average. The size of each structural unit depends on the gelation and drying conditions and varies between 1 to 100 nm in diameter. Correspondingly, pore sizes in an aerogel range from 1 to 100 nm. For silica aerogels skeletal (ie, the solid phase) densities, ρ_s , of 1700–2100 kg/m³ have been reported, whereas the overall densities, ρ , are in the range of 3–500 kg/m³ (54). The fact that the overall density is much lower than the skeletal density suggests that an aerogel is full of open space, or pores. Quantitatively the porosity, P , is given by

$$P = 1 - (\rho/\rho_s)$$

Thus, the porosity of an aerogel is in excess of 90% and can be as high as 99.9%. As a consequence of such a high porosity, aerogels have large internal surface area and pore volume.

Since the pores in an aerogel are comparable to, or smaller than, the mean free path of molecules at ambient conditions (about 70 nm), gaseous conduction of heat within them is inefficient. Coupled with the fact that solid conduction is suppressed due to the low density, a silica aerogel has a typical thermal conductivity of 0.015 W/(m·K) without evacuation. This value is at least an order of magnitude lower than that of ordinary glass and considerably lower than that of CFC (chlorofluorocarbon)-blown polyurethane foams (54).

The low density and high porosity of aerogels lead to several other unique properties. Sound travels in silica aerogels at a longitudinal velocity of 100–300 m/s (the corresponding value in ordinary glass is about 500 m/s) (56). The low sound velocity and low density combine to give a low acoustic impedance, which is the product of the two quantities. Silica aerogels also have a refractive index that is very close to unity, meaning light can enter and leave a piece of aerogel without appreciable reflective losses and refractive effects. And for porosity larger than 0.7, silica aerogels have a very low dielectric constant (less than 2) (57). Finally, compared to silica glass, silica aerogels have an extremely small Young's modulus (10⁶–10⁷ N/m², several orders of magnitude less than nonporous materials), and thus, can be compressed easily (51).

From a practical viewpoint it is not only the low density of aerogels but also the variability of density over a wide range that offers interesting possibilities. As discussed in "Inorganic Materials," a silica gel can be formed under either acidic or basic conditions. Densities that can be formed under either acidic or basic conditions. Densities that can be obtained with these one-step processes are in the range of 25–500 kg/m³. To produce ultralow-density silica aerogels, a two-step process has been developed to extend the lower limit to 3 kg/m³ (58). In the first step, either tetramethyl- or tetraethyl-orthosilicate is reacted with a substoichiometric amount of water to form a partially hydrolyzed, partially condensed silica precursor. After distilling off the alcohol solvent, this precursor is stored in a nonalcoholic solvent. The second step involves the further hydrolysis of the precursor under basic conditions to form a gel. The ability to vary density over two orders of magnitude is significant because almost all the properties discussed in this section vary with density. Specifically (52,54), the following:

Property	Value
refractive index	$n = 1 + 2.1 \times 10^{-4} \rho$; ρ in kg/m ³
sound velocity	$v_s \propto \rho^\beta$
Young's modulus	$Y \propto \rho^\alpha$
where for silica	$\beta = (\alpha - 1)/2$ $\alpha = 3.6; \beta = 1.3$

Note that scaling exponents depend on preparative conditions.

Table 2 summarizes the key physical properties of silica aerogels. A range of values is given for each property because the exact value is dependent on the preparative conditions and, in particular, on density.

Table 2. Typical Values of Physical Properties of Silica Aerogels

Property	Values
density, kg/m ³	3–500
surface area, m ² /g	800–1000
pore sizes, nm	1–100
pore volume, cm ³ /g	3–9
porosity, %	75–99.9
thermal conductivity, W/(m·K)	0.01–0.02
longitudinal sound velocity, m/s	100–300
acoustic impedance, kg/(m ² ·s)	10 ³ –10 ⁶
dielectric constant	1–2
Young's modulus, N/m ²	10 ⁶ –10 ⁷

Applications

Thermal Insulation. In addition to their low thermal conductivity, as discussed in the section above, silica aerogels can be prepared to be highly transparent in the visible spectrum region. Thus, they are promising materials as superinsulating window-spacer. To take further advantage of its

high solar transmission, a silica aerogel layer sandwiched between two glass panes can be used to collect solar energy passively. Figure 6 shows a specific arrangement in which an insulating layer of transparent silica aerogel is placed in front of a brick wall, the surface of which is painted black. Solar radiation collected at the black surface is mostly transferred as heat into the house because heat loss through the aerogel layer is minimal. A shade is put into place to prevent overheating if necessary (53). The same principle applies when silica aerogels are used as insulating covers for solar panels.

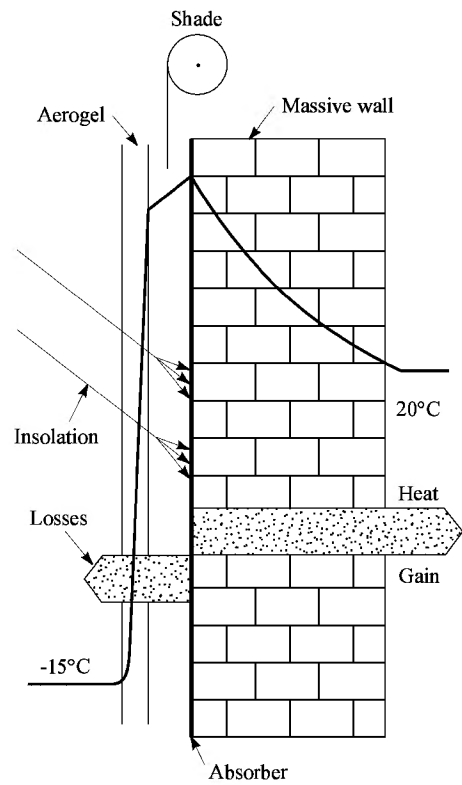


Fig. 6. Schematic diagram showing the use of transparent silica aerogel in passive solar collection. Reproduced from Ref. 53.

Courtesy of Elsevier Science-NL.

The thermal conductivity of silica aerogels can be further reduced by minimizing the radiation leakage with an opacifier such as carbon black (54). The introduction of an opacifier makes the material opaque and unsuitable for window insulation. On the other hand, opaque silica aerogels can be used as insulating materials in appliances such as refrigerators and freezers. Compared with CFC-blown insulating foams, which could release chlorine into the atmosphere, silica aerogels pose no such environmental hazard and are nonflammable. Moreover, the thermal conductivity of opaque silica aerogels is a weak function of temperature (52), making them useful as insulating materials for heat-storage systems. Another promising application of silica aerogels is as a filler in vacuum panel because they do not require a high vacuum for good thermal performance. The high surface area of these materials further allow them to act as a "getter" by adsorbing gases in the panel.

The commercial viability of silica aerogels as thermal insulators depends on the ability to produce them at a competitive price. After all, in the 1950s, the production of Monsanto's Santocel stopped after a lower-cost process to manufacture fumed silica was developed (59). Recently an initial economic analysis that consists of six factors in the manufacturing of aerogels: starting material, solvent, energy, wage, equipment, and facility was published (60). The results show that the dominant cost is the cost of the starting material and that aerogels could be competitive with commercial insulating materials on a cost per R value basis. Indeed, BASF has developed a silica aerogel which has the registered trademark Basogel (61). Since supercritical drying, even with carbon dioxide at a lower temperature, is an energy-intensive process, NanoPore, Inc. is developing an ambient approach to make silica aerogels (49,59). Technological progresses in the next several years will be critical in determining whether aerogels can capture a significant share of the commercial insulation market, which is probably their largest potential area of application. At least two U.S. companies are currently developing aerogels as insulating materials. Aspen Systems manufactures silica aerogels in the forms of powders, monoliths, and blankets. Their present (1996) price range is from \$100 to \$2,000 per cubic foot, depending on the size of the order (62). Aerojet Corporation has collaborated with different end-users in evaluating the market potential of organic aerogels (59), which have even lower thermal conductivities than their silica counterparts.

Catalysis. Kistler explored the catalytic applications of aerogels in the 1930s because of the unique pore characteristics of aerogels (24), but this area of research stayed dormant for about three decades until less tedious procedures to produce the materials were introduced (25,26). Three recent review articles summarize the flurry of research activities since then (63–65). Table 3 is a much abbreviated list of what has been cited in these three articles to demonstrate simply the wide range of catalytic materials and reactions that have been studied.

Table 3. Example of Materials and Reactions Involving Catalytic Aerogels

Materials	Reactions	Examples
Type I, simple oxides		silica, alumina, titania, zirconia
Type II, mixed oxides		nickel oxide–alumina, titania–silica, zirconia–silica, chromia–alumina
Type III, ternary oxides		nickel–oxide–silica–alumina, magnesia–alumina–silica, titania–silica–vanadia, alumina–chromia–thoria
Type IV, supported metals		palladium–alumina, ruthenium–silica, platinum–titania, palladium–silica–alumina
Type V, doped oxides (dopant not an oxide)		zirconia–sulfate, zirconia–phosphate, niobia–phosphate, TiCl ₄ –alumina
Types I, II, III, IV	partial oxidation	isobutylene to methacrolein, acetaldehyde to acetic acid
Type IV	hydrogenation	cyclopentadiene to cyclopentene, benzene to cyclohexane
Types I, II, V	isomerization	1-butene to <i>cis</i> - and <i>trans</i> -2-butene, <i>n</i> -butane to isobutane
Type II	epoxidation	1-hexene and cyclohexene to the corresponding epoxides
Type IV	hydrotreating	hydrodenitrogenation and hydrodesulfurization
Type V	polymerization	ethylene to polyethylene

Most of the studies on catalytic aerogels reported in the open literature involve testing powder samples in experimental reactors. For the potential scale-up to commercial operations, aerogels have several limitations which, ironically, arise from their unique properties. First, even though aerogels have high specific surface area, they also have low densities. The product of the two quantities, which is surface area per unit volume, does not offer a significant advantage over other materials that have been used as adsorbents or catalysts. Second, the low thermal conductivity of aerogels means that it would be difficult to transfer heat in or out of a catalytic packed bed. Third, aerogels are fragile and do not withstand mechanical stress well. Attempts have been made to overcome the last two limitations by supporting or combining aerogels with materials that are either more thermally conducting or more rigid (63). Finally, as in the case of thermal insulation, catalytic aerogels need to be cost-competitive, even though the economic pressure is not as severe because the cost of a catalyst is usually a small fraction of the value of its derived products.

It is commonly believed that catalytic aerogels are interesting because their composition, morphology, and structure can be controlled at a microscopic level. But this feature is generic to any sol-gel-derived materials, so perhaps what set aerogels apart is that during solvent removal, there is another level control over the physical and chemical properties of the products. For example, supercritically dried materials are usually mesoporous and, as such, should be good catalysts for liquid-phase reactions for which there could be diffusional limitations. And for multicomponent materials, the distribution (or mixing state) of the various components could be better preserved without compromising the integrity of the porous network. A nice illustration of these effects is the recent results on epoxidation (66–68).

For the epoxidation of olefins over catalysts containing titania and silica, two factors are considered to be crucial: the accessibility of large pores for the bulky olefins (containing six to ten carbons) and the presence of Ti–O–Si linkages. It has been shown (66–68) that both of these desirable properties can be obtained only by supercritically drying an optimized gel at low temperatures with carbon dioxide. High temperature supercritical drying led to the segregation of titania and destruction of Ti–O–Si linkages. Conventional evaporative drying maintained the density of Ti–O–Si linkages but resulted in microporous xerogels that were inactive. Figure 7 (69) illustrates these observations for the epoxidation of cyclohexene. Besides reporting similar results for other olefins such as cyclododecene, norbornene, and limonene, these researchers were able to establish a semiquantitative correlation between activity and Ti–O–Si connectivity.

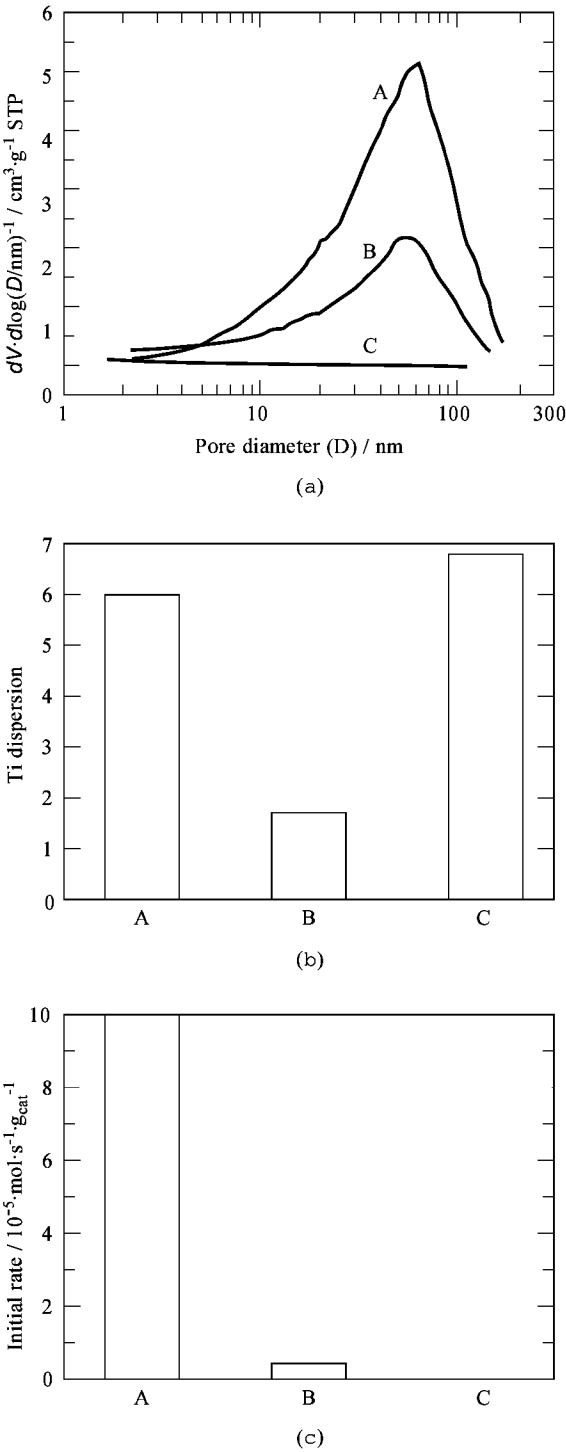


Fig. 7. The effect of preparation on the pore size distribution (a), titanium dispersion (b), and the activity for epoxidation of cyclohexene (c) of titania-silica containing 10 wt % titania and calcined in air at 673 K. Sample A, low-temperature aerogel; Sample B, high-temperature aerogel; Sample C, aerogel.

Reproduced from Ref. 69.

Courtesy of Marvel Dekker, Inc.

The Ti–O–Si linkages in titania–silica also have a large impact on the acidic properties of this mixed oxide. It was recently demonstrated that, over the entire composition range, the extent of mixing as controlled by prehydrolysis changes the acid site density, acid site type (Lewis versus Brønsted), and 1-butene isomerization activity of titania–silica aerogels (70). In fact, these observations appear to be general for other mixed oxide pairs in that sol–gel chemistry affords a higher degree of control over the intimacy of molecular-scale mixing that is not available by other preparative methods (71). Intimate mixing is actually undesirable in some instances, as shown for the selective catalytic reduction of NO with NH₃ (72). For this system, titania crystallites are believed to be good for stabilizing a two-dimensional overlayer of vanadia. Still, for this reaction the role of oxide–oxide interactions in affecting surface acidic properties remains critical (73). The fact that multicomponent materials of specific pore characteristics can be prepared to be either well-mixed or poorly-mixed, depending on the application, represents a unique advantage of catalytic aerogels.

Scientific Research. There are some applications that require such specific and unique properties that an aerogel is the only alternative. In these situations price is no longer a factor and aerogels have been and will continue to be produced for scientific purposes. One example is the use of silica aerogel monoliths as radiators in Chernikov counters, or detectors, in high energy physics and nuclear astrophysics experiments (23). In order to measure precisely the momentum of elementary particles, which are produced by particle interactions in high energy accelerators, over different momentum ranges, it is necessary to vary the refractive index of radiators. The reason is that a charged particle produces light in passing through a medium only if its velocity is higher than the velocity of light in that medium (52). As pointed out in the section "Properties," the refractive index of silica aerogels is related to density. Thus, by varying the density from about 50 to 300 kg m⁻³, which is easily attainable with the control of sol–gel chemistry, materials of indices of refraction between 1.01 and 1.06 can be delivered to measure low values of momentum. Indeed, two large blocks of silica aerogels have been prepared as particle detectors for use in the CERN Intersecting Storage Rings and the Deutsches Elektronen-Synchrotron (DESY) (23).

Aerogels are also among the few materials that can capture cosmic particles intact (74). When a spacecraft approaches a source (eg, the corona of a comet), the cosmic particles have relative speeds of the order of tens of km s⁻¹. These hypervelocity particles tend to melt or vaporize upon collision with a solid. Mesoporous materials that are of very low density can capture these particles relatively intact, but a silica aerogel is the only one that allows the captured particles to be located easily because of its transparency. In fact, as a particle enters a silica aerogel at hypervelocity, it creates a carrot-shaped tract that points to where the particle stops. The high surface area of an aerogel provides the added advantage of adsorbing any volatiles that are either in space or generated during capture.

Panels of silica aerogels have already been flown on several Space Shuttle missions (74). Currently a STARDUST mission has been planned by NASA to use aerogels to capture cometary samples (>1000 particles of >15 micron diameter) and interstellar dust particles (>100 particles of >0.1 micron diameter). The mission involves the launch of a spacecraft in February 1999 to encounter the comet Wild 2 in December 2003 and to return the captured samples to earth in January 2006. Silica aerogels of ultralow density, high uniformity, and high purity need to be prepared to meet the primary objectives of the mission. Furthermore, there are challenges to produce nonsilicate aerogels of the same high quality and aerogels that can perform chemical as well as physical capture.

Besides being used as a tool for scientific research, silica aerogels can be the cause for new scientific phenomena. For example, the long-range correlations of the disorder in silica aerogels are believed to be responsible for the intriguing observations of the superfluid transitions in ⁴He and ³He and on the ordering of ³He–⁴He mixtures (75).

Other Applications. There are several other applications that take advantage of the unique properties of silica aerogels shown in Table 2. In a piezoceramic ultrasound transducer, the signal is reduced as sound waves cross the interface of two materials that have very different acoustic impedances (eg, the acoustic impedances of air and piezoelectric ceramics differ by several orders of magnitude). With its tunable acoustic impedance, a layer of silica aerogel sandwiched between a piezoceramic and air minimizes the mismatch and enhances the signal (54). The low dielectric constant of silica aerogels suggests their usefulness as layer materials in integrated circuits. For silica aerogels to replace the currently used silica glasses, several issues such as the control of porosity, mechanical strength, thermal stability, and process integration problems need to be addressed (57). Finally, silica aerogels are effective dehydrating agents because of their high capacity for absorbing moisture and chemical inertness. For household insect control, silica aerogels work by absorbing waxes from the cuticle of insects, which then die from dehydration; their potency can be enhanced with the doping of insecticides (76).

The introduction of organic aerogels has led to interesting areas of applications. Recent data (77,78) showing the thermal properties of these materials and, more importantly, how these properties can be varied with preparative conditions, should help their further development as thermal insulators. Carbon aerogels are also promising electrode materials in electrochemical double-layer capacitors. Large specific capacities and capacitance densities have been demonstrated in a device consisting of two carbon aerogel wafers; this performance is attributed to the high specific surface area and contiguous structure of the carbon electrode (79). These results are encouraging for the use of carbon aerogels in building low cost, high power, and high energy density capacitors (often referred to as "supercapacitors"). Another potential application of carbon–aerogel electrodes is the capacitive deionization of water. In a recently developed process, water with various anions and cations is pumped through an electrochemical cell consisting of a stack of carbon aerogel electrodes (80). The large specific surface area of the electrodes allows the electrostatic capture of ions. After water is purified, the cell can be regenerated by a simple electrical discharge.

Summary

As of 1996 the commercial market for aerogels remained very small. However, aerogels have the potential of being marketable both as a commodity chemical (eg, in thermal insulation) and as a specialty chemical (eg, in electronic applications) because of their unique and tailorable properties. The next few years will be critical in assessing whether aerogels can penetrate and grow in either end of the market, as the field is changing rapidly with the development of cost-competitive technologies and novel applications.

BIBLIOGRAPHY

1.

S. S. Kistler, *Nature*, **127**, 741 (1931).

2.

C. J. Brinker and G. W. Scherer, *Sol-Gel Science: The Physics and Chemistry of Sol-Gel Processing*, Academic Press, New York, 1990.

3.

J. Livage, M. Henry, and C. Sanchez, *Prog. Solid State Chem.* **18**, 259 (1988).

4.

R. C. Mehrota, *J. Non-Cryst. Solids* **145**, 1 (1992).

5.

H. D. Gesser and P. C. Goswami, *Chem. Rev.* **89**, 765 (1989).

6.

R. K. Iler, *The Chemistry of Silica*, John Wiley & Sons, New York, 1979, Chapt. 3.

7.

J. Y. Ying, J. B. Benziger, and A. Navrotsky, *J. Am. Ceram. Soc.* **76**, 2571 (1993).

8.

B. Handy, K. L. Walther, A. Wokaun, and A. Baiker, in P. A. Jacobs, P. Grange, and B. Delmon, eds., *Preparation of Catalysts V*, Elsevier, Netherlands, 1991, pp. 239–246.

9.

E. I. Ko, in G. Ertl, H. Knözinger, and J. Weitkamp, eds., *Handbook of Heterogeneous Catalysts*, VCH, Germany, 1997, Vol. 1, Sec. 2.1.4.

10.

B. E. Yoldas, *J. Mater. Sci.* **14**, 1843 (1979).

11.

J. Livage and C. Sanchez, *J. Non-Cryst. Solids* **145**, 11 (1992).

12.

U.S. Pat. 4,873,218 (Oct. 10, 1989); **U.S. Pat. 4,997,804** (Mar. 5, 1991), R. W. Pekala.

13.

R. W. Pekala and C. T. Alviso, *Mat. Res. Soc. Symp. Proc.* **270**, 3 (1992).

14.

R. W. Pekala and C. T. Alviso, *Mat. Res. Soc. Symp. Proc.* **180**, 791 (1990).

15. R. W. Pekala, C. T. Alviso, X. Lu, J. Gross, and J. Fricke, *J. Non-Cryst. Solids* **188**, 34 (1995).
16. B. M. Novak, *Adv. Mater.* **5**, 422 (1993).
17. C. Sanchez and F. Ribot, *New J. Chem.* **18**, 1007 (1994).
18. D. Avnir, *Acc. Chem. Res.* **28**, 328 (1995).
19. U. Schubert, N. Häsing, and A. Lorenz, *Chem. Mater.* **7**, 2010 (1995).
20. D. A. Loy and K. J. Shea, *Chem. Rev.* **95**, 1431 (1995).
21. D. Avnir, S. Braun, O. Lev, and M. Ottolenghi, *Chem. Mater.* **6**, 1605 (1994).
22. W. Cao, X. Y. Song, and A. J. Hunt, *Mat. Res. Soc. Symp. Proc.* **349**, 87 (1994).
23. A. J. Ayen and P. A. Iacobucci, *Rev. Chem. Eng.* **5**, 157 (1988).
24. S. S. Kistler, S. Swan, Jr., and E. G. Appel, *Ind. Eng. Chem.* **26**, 388 (1934).
25. S. J. Teichner, *CHEMTECH*, **21**, 372 (1991).
26. S. J. Teichner, G. A. Nicolaon, M. A. Vicarini, and G. E. E. Gardes, *Adv. Colloid Interface Sci.* **5**, 245 (1976).
27. **U.S. Pat. 4,610,863** (Sept. 9, 1986), P. H. Tewari and A. J. Hunt.
28. **U.S. Pat. 4,619,908** (Oct. 28, 1986), C. P. Cheng, P. A. Iacobucci, and E. N. Walsh.
29. P. Wang, A. Emmerling, W. Tappert, O. Spormann, J. Fricke, and H. G. Haubold, *J. Appl. Cryst.* **24**, 777 (1991).
30. D. R. Ulrich, *J. Non-Cryst. Solids* **100**, 174 (1988).
31. D. M. Smith, R. Desphande, and C. J. Brinker, *Mat. Res. Symp. Proc.* **271**, 553 (1992).
32. G. Cogliati, M. Guglielmi, T. M. Che, and T. J. Clark, *Mat. Res. Symp. Proc.* **180**, 329 (1990).
33. M. Beghi, P. Chiuro, L. Costa, M. Palladina, and M. F. Pirini, *J. Non-Cryst. Solids* **145**, 175 (1992).
34. C. J. Brodsky and E. I. Ko, *J. Mater. Chem.* **4**, 651 (1994).
35. D. C. M. Dutoit, M. Schneider, and A. Baiker, *J. Porous Mater.* **1**, 165 (1995).
36. M. Schneider and A. Baiker, *Catal. Rev.-Sci. Eng.* **37**(4), 529 (1995).
37. G. W. Scherer, in A. S. Mujumdar, ed., *Drying*, Elsevier, the Netherlands, 1992, pp. 92–113.
38. G. W. Scherer, *J. Non-Cryst. Solids*, **145**, 33 (1992).
39. G. W. Scherer, *J. Sol-Gel Sci. Tech.* **3**, 127 (1994).
40. T. Woignier, G. W. Scherer, and A. Alaoue, *J. Sol-Gel Sci. Tech.*, **3**, 141 (1994).
41. L. W. Hrubesh and J. F. Poco, *J. Non-Cryst. Solids*, **188**, 46 (1995).
42. J. F. Poco, P. R. Coronado, R. W. Pekala, and L. W. Hrubesh, *Mat. Res. Soc. Symp. Proc.* **431**, 297 (1996).
43. R. Deshpande, D. M. Smith, and C. J. Brinker, *Mat. Res. Soc. Symp. Proc.* **271**, 553 (1992).
44. C. J. Brinker, R. Sehgal, N. K. Raman, S. S. Prakash, and L. Delattre, *Mat. Res. Soc. Symp. Proc.* **368**, 329 (1995).
45. D. M. Smith, R. Deshpande, and C. J. Brinker, *Mat. Res. Soc. Symp. Proc.* **271**, 567 (1992).
46. D. M. Smith, R. Deshpande, and C. J. Brinker, *Ceram. Trans.* **31**, 71 (1993).
47. S. S. Prakash, C. J. Brinker, A. J. Hurd, and S. M. Rao, *Nature* **374**, 439 (1995).
48. S. S. Prakash, C. J. Brinker, and A. J. Hurd, *J. Non-Cryst. Solids* **190**, 264 (1995).
49. D. M. Smith, W. C. Akerman, R. Roth, A. Zimmerman, and F. Schwertfeger, *Mat. Res. Soc. Symp. Proc.* **431**, 291 (1996).
50. D. Klvana, J. Chaouki, M. Repellin-Lacroix, and G. M. Pajonk, *J. Phys. Colloq.*, **C4**, 29 (1989).
51. H. D. Gesser and P. C. Goswami, *Chem. Rev.* **89**, 765 (1989).
52. J. Fricke and A. Emmerling, *J. Am. Ceram. Soc.* **75**(8), 2027 (1992).
53. J. Fricke, *J. Non-Cryst. Solids* **147/148**, 356 (1992).
54. J. Fricke and J. Gross, *Mater. Eng.* **8**, 311 (1994).
55. T. Heinrich, U. Klett, and J. Fricke, *J. Porous Mater.* **1**, 7 (1995).
56. J. Fricke, *Sci. Amer.* **256**(5), 92 (1988).
57. D. M. Smith, J. Anderson, C. C. Cho, G. P. Johnston, and S. P. Jeng, *Mat. Res. Soc. Symp. Proc.* **381**, 261 (1995).
58. T. M. Tillotson and L. W. Hrubesh, *J. Non-Cryst. Solids* **145**, 45 (1992).
59. C. M. Caruana, *Chem. Engr. Prog.* (June 11, 1995).
60. G. Carlson, D. Lewis, K. McKinley, J. Richardson, and T. Tillotson, *J. Non-Cryst. Solids* **186**, 372 (1995).
61. G. Herrmann, R. Iden, M. Mielke, F. Teich, and B. Ziegler, *J. Non-Cryst. Solids* **186**, 380 (1995).
62. <http://www.aspensystems.com/aerogel.html>
63. G. M. Pajonk, *Appl. Catal.* **72**, 217 (1991).
64. G. M. Pajonk, *Catal. Today*, **35**, 319 (1997).
65. M. Schneider and A. Baiker, *Catal. Rev.-Sci. Eng.* **37**(4), 515 (1995).
66. D. Dutoit, M. Schneider, and A. Baiker, *J. Catal.* **153**, 165 (1995).
67. R. Hutter, T. Mallat, and A. Baiker, *J. Catal.* **153**, 177 (1995).
68. R. Hutter, T. Mallat, and A. Baiker, *J. Catal.* **157**, 665 (1995).
69. From Ref. 65, p. 544.
70. J. B. Miller, S. T. Johnston, and E. I. Ko, *J. Catal.* **150**, 311 (1994).
71. J. B. Miller and E. I. Ko, *Catal. Today*, **35**, 269 (1997).
72. B. E. Handy, A. Baiker, M. Schramal-Marth, and A. Wokaun, *J. Catal.* **133**, 1 (1992).
73. R. J. Willey, C. T. Wang, and J. B. Peri, *J. Non-Cryst. Solids* **186**, 408 (1995).
74. P. Tsou, *J. Non-Cryst. Solids* **186**, 415 (1995).
75. M. Chan, N. Mulders, and J. Reppy, *Physics Today* (Aug. 30, 1996).
76. <http://www.ent.aggri.umn.edu/academics/classes/ipm/chapters/ware.htm>
77. X. Lu, R. Caps, J. Fricke, C. T. Alviso, and R. W. Pekala, *J. Non-Cryst. Solids* **188**, 226 (1995).
78. V. Bock, O. Nilsson, J. Blumm, J. Fricke, *J. Non-Cryst. Solids* **185**, 233 (1995).
79. S. T. Mayer, R. W. Pekala, and J. K. Kaschmitter, *J. Electrochem. Soc.* **140**(2), 446 (1993).
80. J. C. Farmer, D. V. Fix, G. V. Mack, R. W. Pekala, and J. F. Poco, *J. Electrochem. Soc.* **143**(1), 159 (1996).

Edmond I. Ko
Carnegie Mellon University

AQUACULTURE

One definition of aquaculture is the rearing of aquatic organisms under controlled or semicontrolled conditions (1). Another, promulgated by the Food and Agriculture Organization (FAO) of the United Nations, is that aquaculture is, "the farming of aquatic organisms, including fish, molluscs, crustaceans, and aquatic plants" (2). Included within those broad definitions are activities in fresh, brackish, marine, and even hypersaline waters. The term *mariculture* is often used in conjunction with aquaculture in the marine environment.

Public sector aquaculture involves production of aquatic animals to augment or establish recreational and commercial fisheries. Public sector aquaculture is widely practiced in North America and to a lesser extent in other parts of the world. The FAO definition of aquaculture also indicates that farming implies ownership of the organisms being cultured, which would seem to exclude public sector aquaculture.

In recent years, aquaculture has been increasingly used as a means of aiding in the recovery of threatened and endangered species. Those efforts are currently public sector activities, although there is interest in the private sector to become involved. As global awareness of endangered species issues grows, recovery programs for aquatic threatened and endangered species may arise in many more countries. Going hand in hand with attempts to recover endangered species are enhancement stocking programs aimed at releasing juvenile animals to build back stocks of aquatic animals that have been reduced due to overfishing. Examples of enhancement programs currently in existence include the stocking of cod in Norway, flounders in Japan, and red drum in the United States.

The bulk of global production from aquaculture is utilized directly as human food, with public aquaculture playing a minor role in many nations or being absent. Private aquaculture is not only about human food production, however. There is, in some regions, well-developed private sector aquaculture involved in the production of bait and ornamental fishes and invertebrates.

Aquatic plants are cultured in many regions of the world. In fact, aquatic plants, primarily seaweeds, account for nearly 25% of the world's aquaculture production (3). Most of the information available in the literature relates to the production of such aquatic animals as molluscs, crustaceans, and finfish.

The origins of aquaculture are rooted in China and may go back some 4000 years. Asia dominates the world in aquaculture production, with China producing over 10 million metric tons and Japan and India each producing well over 1 million metric tons (4). In North America, the culture of fish began in the mid-nineteenth century and grew rapidly in the public sector after the establishment of the U.S. Fish and Fisheries Commission in 1871 (5). Private aquaculture existed as a minor industry for many decades, coming into prominence beginning in the 1960s. Since then the United States has become one of the leaders in aquaculture research and development, although production, while significant at over 400,000 metric tons by 1992 (4), amounted to only about 2% of the world's total of nearly 19 million metric tons. The United States commercial aquaculture industry is dominated by channel catfish, trout, salmon, minnows, oysters, mussels, clams, and crawfish. A number of other fishes and invertebrates are also being reared. Included are tilapia, striped bass and hybrid striped bass, red drum, goldfish, tropical fishes, and shrimp. In the public sector, hatcheries produce large numbers of such species as salmon, trout, largemouth and small-mouth bass, sunfish, crappie, northern pike, muskellunge, walleye, and catfish for stocking or growout.

Aquaculture production continues to grow annually, but increasing competition for suitable land and water, problems associated with wastewater from aquaculture facilities, disease outbreaks, and potential shortages of animal protein for aquatic animal feeds are having, or may have, negative effects on future growth. New technology, including the application of molecular genetic approaches to improving performance and disease resistance in aquaculture species, along with the development of water reuse (recirculating) systems and the establishment of offshore facilities, may provide the impetus for a resurgence in growth of the industry.

Economics

The production of aquatic animals for recreation, in nations where that type of aquaculture exists, is typically funded through user fees such as fishing licenses that support hatcheries and the personnel to run them. In order for most private aquaculture companies to get started, outside funding is required. Funding may come through banks and other commercial lending sources or from venture capitalists. The high risks associated with aquaculture have made it difficult for many firms to obtain bank loans, although that situation is changing as bankers become more knowledgeable and comfortable with underwriting aquaculture ventures.

A key factor in obtaining funding support for aquaculture is development of a sound business plan. The plan needs to demonstrate that the prospective culturist has identified all costs associated with establishment of the facility and its day-to-day operation. One or more suitable sites should have been identified and the species to be cultured selected before the business plan is submitted. Cost estimates should be verifiable. Having actual bids for a specific task at a specific location; eg, pond construction, well drilling, building construction, and vehicle costs helps strengthen the business plan.

Land costs vary enormously both between and within countries. Compare the cost of coastal land in south Florida where it might be possible to consider rearing shrimp with that of Mississippi farmland suitable for catfish farming. The former might be thousands of dollars for every meter of ocean front, while the latter may be obtained for one or two thousand dollars per hectare.

The amount of land required varies as well, not only as a function of the amount of production that is anticipated, but also on the type of culture system that is used. It may take several hectares of static culture ponds to produce the same biomass of animals as one modest size raceway through which large volumes of water are constantly flowed. Construction costs vary from one location to another. Local labor and fuel costs must be factored into the equation. The experience of contractors in building aquaculture facilities is another factor to be considered.

The need for redundancy in the culture system needs to be assessed. Failure of a well pump that brings up water to supply a static pond system may not be a serious problem in countries where new pumps can be purchased in a nearby town. However, it can be disastrous in developing countries where new pumps and pump parts are often not available, but must be ordered from another country. Several weeks or months may pass before the situation can be remedied unless the culturist maintains a selection of spares.

The business plan needs to provide projections of annual production. Based on those estimates and assumed food conversion rates (food conversion is calculated by determining the amount of feed consumed by the animals for each kilogram of weight gain), an estimate of feed costs can be made. For many aquaculture ventures, between 40 and 50% of the variable costs involved in aquaculture can be attributed to feed.

Aquaculturists may elect to purchase animals for stocking or maintain their own broodstock and hatchery. The decision may rest on such factors as the availability and cost of fry fish, post-larval fish, oyster spat, or other early life history stages in the location selected for the aquaculture venture.

Land purchases and many of the costs associated with facility development can be accomplished with long-term loans of 15 to 30 years. Equipment such as pumps and trucks are usually depreciated over a few years and are funded with shorter-term loans. Operating expenses for such items as feed, chemicals, fuel, utilities, salaries, taxes, and insurance may require periodic short-term loans to keep the business solvent. The projected income should be based on a realistic estimate of farmgate value of the product and an accurate assessment of anticipated production. Each business plan should project income and expenses projected over the term of all loans in order to demonstrate to the lending agency or venture capitalist that there is a high probability the investment will be repaid.

Regulation

The extent to which governments regulate aquaculture varies greatly from one nation to another. In some parts of the world, particularly in developing nations, there has historically been little or no regulation. Inexpensive land and labor, low taxes, excellent climates, and a lack of government interference have drawn many aquaculturists to underdeveloped countries, most of which are in the tropics. Unregulated expansion of aquaculture in some countries has led to pollution problems, destruction of valuable habitats such as mangrove swamps, and has enhanced the spread of disease from one farm to another. The need for imposing regulations is now becoming evident around the world. Response to that need varies considerably from one nation to another.

In developed countries there may or may not be a standardized set of national regulations. The United States is an example of a mixture of local, state, and federal regulations. Permits from a country, state, or federal agency may be required for drilling wells, pumping water, releasing water, use of exotic species, constructing facilities, etc. In the United States, most permits can be obtained at the local or state level. In some instances the federal government has deferred permitting authority to the states when state regulations are as rigorous or more so than national regulations. Federal agencies become involved when aquaculture projects are conducted in navigable waters (U.S. Army Corps of Engineers) or might impact threatened or endangered species (National Biological Service).

In general, it is easier to establish an aquaculture facility on private land than in public waters such as a lake or coastal embayment. Prospective aquaculturists who want to establish facilities in public waters may be confronted at public hearings by outraged citizens who do not want to see an aquaculture facility in what they consider to be their water. The issue is highly contentious in some nations (eg, the United States). In other countries, aquaculture in public waters is seen as not only a good use of natural resources, but can be considered an amenity (eg, Japan).

Obtaining permits is often not simple. Few states have one office that can accommodate the prospective aquaculturist. In most cases it is necessary to contact a number of state agencies to apply for permits. Public hearings may be required before permits are approved. The process can take months or even years to complete. The costs involved in going through the process may be prodigious. After the expenditure of considerable amounts of time and money, there is no guarantee that the permits will ultimately be granted.

Most states now have an aquaculture coordinator, usually housed in the state department of agriculture, who can assist prospective aquaculturists in finding a path through the permitting process. Anyone considering development of an aquaculture facility should become educated on the permitting process of the state or nation in which the facility will be developed. In cases where the process is involved, it should be initiated well in advance of the anticipated time of actual facility construction.

Species Under Cultivation

This article emphasizes aquatic animal production, but many hundreds of thousands of people are involved, worldwide, in aquatic plant production. The quantity of brown seaweeds, red seaweeds, green seaweeds, and other algae produced in 1992 was estimated at over 4.8 million metric tons (Table 1). Miscellaneous aquatic plants such as watercress and water chestnuts contributed an additional 600,000 metric tons. Microscopic algae and cyanobacteria are sometimes marketed as food or as a nutritional supplement (eg, *Spirulina* sp.). In addition, an undocumented quantity of algae (mostly of the single-celled variety) is produced for use as food for filter-feeding aquatic animals (primarily molluscs and zooplankton). Planktonic organisms such as rotifers are reared on algae and then used to feed the young of crustaceans and fishes that do not accept prepared feeds.

Table 1. World Aquaculture Production in 1992 for Selected Groups of Aquaculture Species^a

Species group	Production, 10 ³ t
carp and other cyprinids	6,652
tilapia and other cichlids	473
salmon and trout	628
flatfish	120
freshwater crustaceans	624
shrimp	884
oysters	954
mussels	109
scallops	549
clams, cockles, arkshells	765
brown seaweeds	3,640
red seaweeds	1,133
green seaweeds and other algae	17
miscellaneous aquatic plants	600
<i>Total^b</i>	<i>19,311</i>

^a Ref. 4.
^b Includes species categories not listed.

Animal aquaculture is concentrated on finfish, molluscs, and crustaceans. Sponges, echinoderms, tunicates, turtles, frogs, and alligators are being cultured, but production is insignificant in comparison with the three principal groups. Common and scientific names of many of the species of the finfish, molluscs, and crustaceans currently under culture are presented in Table 2. Included are examples of bait, recreational, and food animals.

Table 2. Examples of Common and Scientific Names of Selected Aquaculture Species

Type of organism	Common name	Scientific name
finfish	African catfish	<i>Clarias gariepinus</i>
	Atlantic halibut	<i>Hippoglossus hippoglossus</i>
	Atlantic salmon	<i>Salmo salar</i>
	Bighead carp	<i>Aristichthys nobilis</i>
	Bigmouth buffalo	<i>Ictiobus bubalus</i>
	Black crappie	<i>Pomoxis nigromaculatus</i>
	Blue catfish	<i>Ictalurus furcatus</i>
	Blue tilapia	<i>Tilapia aurea</i>
	Bluegill	<i>Lepomis macrochirus</i>
	Brook trout	<i>Salvelinus fontinalis</i>
	Brown trout	<i>Salmo trutta</i>
	Catla	<i>Catla catla</i>
	Channel catfish	<i>Ictalurus punctatus</i>
	Chinook salmon	<i>Oncorhynchus tshawytscha</i>

	Chum salmon	<i>Oncorhynchus keta</i>
	Coho salmon	<i>Oncorhynchus kisutch</i>
	Common carp	<i>Cyprinus carpio</i>
	Fathead minnow	<i>Pimephales promelas</i>
	Gilthead sea bream	<i>Sparus aurata</i>
	Goldfish	<i>Carassius auratus</i>
	Grass carp	<i>Ctenopharyngodon idella</i>
	Largemouth bass	<i>Micropterus salmoides</i>
	Milkfish	<i>Chanos chanos</i>
	Mossambique tilapia	<i>Tilapia mossambica</i>
	Mrigal	<i>Cirrhinus mrigala</i>
	Mud carp	<i>Cirrhina molitorella</i>
	Muskellunge	<i>Esox masquinongy</i>
	Nile tilapia	<i>Tilapia nilotica</i>
	Northern pike	<i>Esox lucius</i>
	Pacu	<i>Colossoma metrei</i>
	Pink salmon	<i>Oncorhynchus gorbuscha</i>
	Plaice	<i>Pleuronectes platessa</i>
	Rabbitfish	<i>Siganus</i> spp.
	Rainbow trout	<i>Oncorhynchus mykiss</i>
	Red drum	<i>Sciaenops ocellatus</i>
	Rohu	<i>Labeo rohita</i>
	Sea bass	<i>Dicentrarchus labrax</i>
	Shiners	<i>Notropis</i> spp.
	Silver carp	<i>Hypophthalmichthys molitrix</i>
	Smallmouth bass	<i>Micropterus dolomieu</i>
	Sole	<i>Solea solea</i>
	Steelhead	<i>Oncorhynchus mykiss</i>
	Striped bass	<i>Morone saxatilis</i>
	Walking catfish	<i>Clarias batrachus</i>
	Walleye	<i>Stizostedion vitreum vitreum</i>
	White crappie	<i>Pomoxis annularis</i>
	Yellow perch	<i>Perca flavescens</i>
	Yellowtail	<i>Seriola quinqueradiata</i>
molluscs	American oyster	<i>Crassostrea virginica</i>
	Bay scallop	<i>Aequipecten irradians</i>
	Blue mussel	<i>Mytilus edulis</i>
	Northern quahog	<i>Mercenaria mercenaria</i>
	Pacific oyster	<i>Crassostrea gigas</i>
crustaceans	Southern quahog	<i>Mercenaria campechiensis</i>
	Freshwater shrimp	<i>Macrobrachium rosenbergii</i>
	Blue shrimp	<i>Penaeus stylirostris</i>
	Kuruma shrimp	<i>Penaeus japonicus</i>
	Pacific white shrimp	<i>Penaeus vannamei</i>
	Red swamp crawfish	<i>Procambarus clarkii</i>
	Tiger shrimp	<i>Penaeus monodon</i>
	White river crawfish	<i>Procambarus acutus acutus</i>
	White shrimp	<i>Penaeus setiferus</i>
	California giant kelp	<i>Macrocystis pyrifera</i>
algae (seaweeds)	Eucheuma	<i>Eucheuma cottoni</i>
	False Irish moss	<i>Gigartina stellata</i>
	Gracilaria	<i>Gracilaria</i> sp.
	Irish moss	<i>Chondrus crispus</i>
	Laminaria	<i>Laminaria</i> spp.
	Nori or laver	<i>Porphyra</i> spp.
	Wakame	<i>Undaria</i> spp.

Various species of carp and other members of the family *Cyprinidae* lead the world in terms of quantity of animals produced. In 1992 the total was over 6.6 million metric tons (see Table 1). China is the leading carp producing nation, and is the world's leading aquaculture nation overall (Table 3). Significant amounts of carp are also produced in India and parts of Europe.

Table 3. Top Aquaculture Producing Nations in 1992 and Amount of Production in Each ^a

Nation	Production ^b , 10 ³ t
China	10,409
Japan	1,397
India	1,374
Republic of Korea	955
Philippines	737
Indonesia	690
United States	413
Thailand	359

^a Ref. 4.

^b Includes aquatic plants.

Fishes in the family *Salmonidae* (trout and salmon) are in high demand, with the interest in salmon being greatest in developed nations. Salmon, mostly Atlantic salmon, are produced in Canada, Chile, Norway, New Zealand, Scotland, and the United States. Fishes in the family *Cichlidae*, which includes several cultured species of tilapia, are reared primarily in the tropics, but have been widely introduced throughout both the developed and developing world.

Catfish are not a major contributor to aquaculture production globally, but the channel catfish industry dominates United States aquaculture. United States catfish production, primarily channel catfish, was 209,090 metric tons in 1992 (4).

Among the invertebrates, most of the world's production is associated with mussels, oysters, shrimp, scallops, and clams. Crawfish culture is of considerable importance in the United States, but amounted to only 24,211 metric tons in 1992 (4); insignificant compared to some other invertebrate species.

Small amounts of crabs, lobsters, and abalone are being cultured in various nations. All three bring good prices in the marketplace but have drawbacks associated with their culture. Crabs and lobsters are highly cannibalistic. Rearing them separately to keep them from consuming one another during molting has precluded economic culture in nearly every instance. Abalone eat seaweeds and can only be reared in conjunction with a concurrent seaweed culture facility or in regions where large supplies of suitable seaweeds are available from nature. In some instances the value of the seaweed for direct human consumption may make the highest and best use of the plants.

In all, there are perhaps 100 species of aquatic animals under culture. Many researchers have turned their attention to species for which there is demand by consumers, but for which the technology required for commercial production is not available. Examples are dolphin, also known as mahimahi (*Coryphaena hippurus*), Pacific halibut (*Hippoglossus stenolepis*), southern flounder (*Paralichthys lethostigma*), winter flounder (*Pseudopleuronectes americanus*), American lobster (*Homarus americanus*), and blue crab (*Callinectes sapidus*). Each of the species mentioned is marine and has small eggs and larvae. Providing the first feeding stages with acceptable food has been a common problem, as has the fragility of the early life stages of many species, and the problem of cannibalism.

Fishes with large eggs, such as trout, salmon, catfish, and tilapia, were among the first to be economically successful in modern times. Small eggs do not necessarily mean that sophisticated research is required to develop the technology required for successful culture. Carp, which have been cultured in China for millenia, have extremely small eggs. At the time the methodology for carp culture was developed, there were no research scientists, although there must have been dedicated farmers who used their common sense and trial-and-error methods to establish aquaculture.

Culture Systems

At one extreme aquaculture can be conducted with a small amount of intervention from humans and the employment of little technology. At the other is total environmental control and the use of computers, molecular genetics, and complex modern technology. Many aquaculturists operate between the extremes. The range of culture approaches can be described as running from extensive to intensive, or even hyper-intensive, with extensive systems being relatively simple and intensive systems being complex to very complex. In general, as the level of culture intensity increases, stocking density, and as a consequence, production per unit area of culture system or volume of water, increases.

The most extensive types of aquaculture involve minimal human intervention to promote increases in natural productivity. Good examples can be found relative to oyster and pond fish culture. With respect to oysters, one of the most extensive forms of culture involves placing oyster, clam, or other types of shell (cultch) on the bottom in intertidal areas that are known to have good oyster reproduction, but limited natural cultch material. The additional substrate may subsequently be colonized by oyster larvae (spat), thereby potentially increasing productivity. The next level of intensity might involve placing bags of cultch out in nature to collect spat in a productive area that already has sufficient quantities of natural cultch. After spat settlement, the bags of shells would be moved to an area where limited natural substrate availability has led to low productivity. The growout area may be held as a common resource, or it may be made commercially available on a leasehold basis.

The next step in increasing oyster culture intensity might involve hatchery production and settling of spat on cultch. Once again, the cultch would be later distributed over a bed leased or owned by the oyster culturist (Fig. 1). Control of predators such as starfish and oyster drills could easily be a part of culture at all levels.

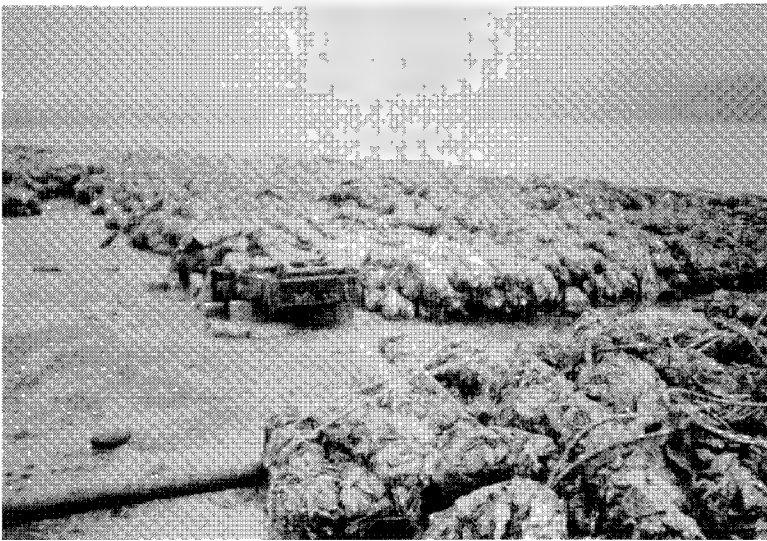


Fig. 1. Bags of oyster shell used as cultch for the settling of oyster spat. The spat-laden shell is ultimately distributed on leased oyster beds.

The highest level of intensity with respect to oysters involves hatchery production of spat and the rearing of them suspended from rafts, long-lines, or as cultchless oysters in trays. In the raft and longline techniques cultch material to which oyster spat or larval mussels are attached is strung on ropes (longlines) suspended from floating rafts or from other ropes held parallel to the water surface with buoys. The lines suspended from the ropes (called strings) of growing oysters are of such a length that the young shellfish are kept within the photic zone and not allowed to touch the bottom where starfish, oyster drills, and other predators can attack. Scallops, which do not attach to cultch, are sometimes grown in bags suspended from long lines or rafts. Similarly, young mussels can be held in proximity to strings with fine mesh materials that retains the animals in place until they attach (by means of what is called a byssus thread) to the string.

The stocking of ponds, lakes, and reservoirs to increase the production of desirable fishes that depend on natural productivity for their food supply and are ultimately captured by recreational fishermen or for subsistence is another example of extensive aquaculture. Some would consider such practices as lying outside of the realm of aquaculture, but since the practice involves human intervention and often employs fishes produced in hatcheries, recreational or subsistence level stocking is associated with, if not a part of aquaculture. Similarly, stocking new ponds or water bodies which have been drained or poisoned to eliminate undesirable species prior to restocking, can lead to increased production of desirable species.

Most of the aquaculture practiced around the world is conducted in ponds (Fig. 2). Ponds range in size but production units are generally 0.1 to 10 ha in area. The intensity of aquaculture in ponds can range from a few kg ha to thousands of kg ha of annual production.



Fig. 2. Aquaculture ponds are often rectangular in shape. They should be equipped with plumbing for both inflow and drainage of water.

Fertilization of ponds to increase productivity is the next level of intensity with respect to fish culture, followed by provision of supplemental feeds. Supplemental feeds are those that provide some additional nutrition but cannot be depended upon to supply all the required nutrients. Provision of complete feeds, those that do provide all of the nutrients required by the fish, translates to another increase in intensity. Associated with one or more of the stages described might be the application of techniques that lead to the maintenance of good water quality. Examples are continuous water exchange, mechanical aeration, and the use of various chemicals used to adjust such factors as pH, alkalinity, and hardness.

With the application of increased technology and control over the culture system, intensity continues to increase. Utilization of specific pathogen-free animals, provision of nutritionally complete feeds, careful monitoring and control of water quality, and the use of animals bred for good performance, can lead to impressive production levels. The United States channel catfish industry is a good example. During the early 1960s pond production levels were typically about 1500 kg/ha/yr. As better feeds and management practices were developed, production increased to an average of about 3000 kg/ha/yr by the next decade. In the 1980s, nutritionally complete feeds were perfected, better prediction and amelioration of water quality problems had been developed, and diseases were being better avoided or controlled. Catfish farmers typically produced 4000 kg/ha/yr and some, who used aeration and exchanged water during part of the growing season, were able to produce 10,000 kg/ha/yr or more.

Where water is plentiful and inexpensive, raceway culture is an attractive option and one which allows for production levels well in excess of what is possible in ponds. Trout are frequently reared in linear raceways from hatching to market size. Linear raceways are longer than they are wide, and are usually no deeper than 1–2 m (Fig. 3). High density raceways used in production facilities are commonly constructed of poured concrete. Small raceways of the type used in hatcheries and research facilities may be constructed of fiber glass or other resilient materials. Water is introduced at one end and flows by gravity through the raceway to exit the other end. Circular raceways, called tanks (Fig. 4), are also used by aquaculturists. Tanks are usually no more than 2 m deep and may be from less than 1 m to as much as 10 m in diameter. Concrete tanks can be found, but most are constructed of fiber glass, metal, or wood that is sealed and covered with epoxy or some other waterproof material. Plastic liners are commonly used in metal or wood tanks to prevent leakage, and in the case of metal, to avoid exposing the aquaculture animals to trace element toxicity.

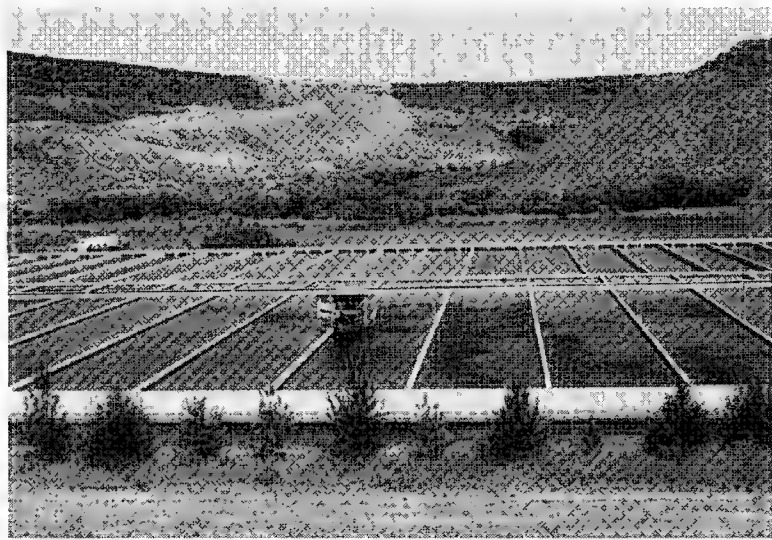


Fig. 3. A commercial trout facility in Idaho, U.S. Linear raceways are commonly used for the production of trout from fry to either release or market and for salmon from fry to smolt size.



Fig. 4. Circular tanks are a raceway option. They can be placed outdoors or used in conjunction with indoor water systems. Circular tanks are commonly used in recirculating systems.

Linear raceways are commonly used by trout and salmon culturists both for commercial production and for hatchery programs conducted by government agencies. Large numbers of state and federal salmon hatcheries in Washington and Oregon, along with governmental and private hatcheries in Alaska, collect and fertilize eggs, hatch them, and rear the young fish to the smolt stage at which time they become physiologically adapted to enter seawater.

Commercial salmon culturists can rear their fish to market size in freshwater raceways although most salmon are grown from smolt to market size or adulthood in the marine environment, either as free roaming fish or in confinement. Since salmon instinctively return to spawn in the waters where they were hatched, it is possible to establish hatcheries and smolt-rearing facilities that take advantage of the homing instinct. The technique is known as ocean ranching. When the fish that had been released as smolts return to spawn, sufficient numbers of adults are collected for use as broodfish to continue the cycle. The remainder may be harvested by the aquaculturist or by commercial fishermen after which the fish are processed and marketed.

Salmon, steelhead trout, and a variety of marine fishes are currently being reared in net-pens (Fig. 5). The typical salmon net-pen is several meters on each side and may be as much as 10 m deep (1). Smaller units, called cages, are sometimes used by freshwater culturists. Cages tend to have volumes of no more than a few cubic meters.



Fig. 5. Marine net-pens such as the ones shown here in Puget Sound, Washington (U.S.), are used for the rearing of salmon by commercial fish farmers.

Net-pen technology was developed in the 1960s, but has only been widely employed commercially for salmon production since the 1980s when the Norwegian salmon farming industry was developed. The Japanese began producing large numbers of sea bream and yellowtail in net-pens during the 1960s. Other nations have employed the technology as well. Most net-pens are located in protected waters since they are easily damaged or destroyed by storms.

Competition by various user groups for space in protected coastal waters in much of the world has led to strict controls and in some cases prohibitions against the establishment of inshore net-pen facilities. As a result, there is growing interest in developing the technology to move offshore. Various designs for offshore net-pens have been developed and a few have been tested (Fig. 6). A number of different designs, including systems that are semi- or totally submersible, have been able to withstand storm waves of at least 6 m, but the costs of those systems are very high compared with inshore net-pens, so commercial viability has yet to be demonstrated.

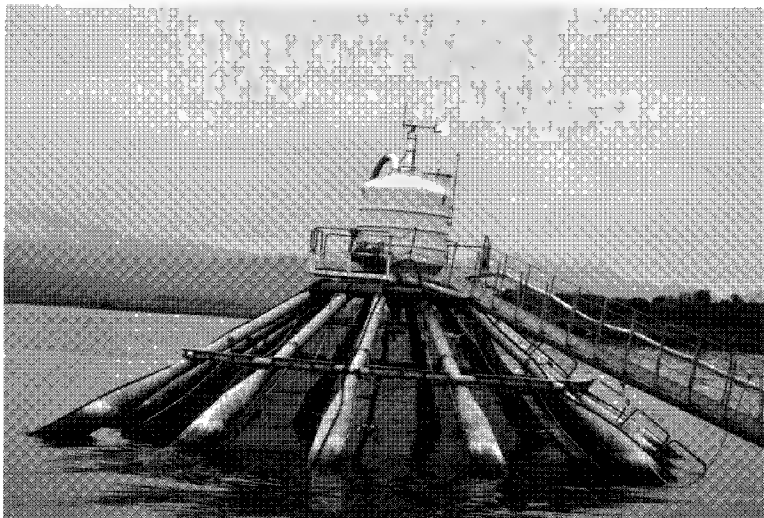


Fig. 6. A salmon net-pen in Scotland designed for use offshore, and in this case, exposed coastal waters.

The highest levels of intensity that can be found in aquaculture systems are associated with totally closed systems, often called recirculating systems. In these systems, all water passing through the chambers in which the finfish or shellfish are held is continuously treated and reused. Once filled initially, closed systems can theoretically be operated for long periods of time without water replacement. It is necessary to add some water to such systems to make up for that lost to evaporation, splashout, and in conjunction with solids removal. Many of the recirculating systems in use today are operated in a mode between entirely closed and completely open. In most there is a significant percentage of replacement water added either continuously or intermittently on a daily basis. Such partial recirculating systems may exchange from a few percent to several hundred percent of system volume each day.

The heart of a recirculating water system is the biofilter, a device that contains a medium on which bacteria that help purify the water become established (Fig. 7). Fish and aquatic invertebrates produce ammonia as a primary metabolite. If not removed or converted to a less toxic chemical, ammonia can quickly reach lethal levels. Two genera of bacteria are responsible for ammonia removal in biofilters. The first, *Nitrosomonas*, converts ammonia (NH_3) to nitrite (NO_2^-). The second, *Nitrobacter*, converts nitrite to nitrate (NO_3^-). Nitrite is highly toxic to aquatic animals, although nitrate can be allowed to accumulate to relatively high levels. If both genera of bacteria are active, the conversion from ammonia through nitrite to nitrate is so rapid that nitrite levels remain within the safe range.

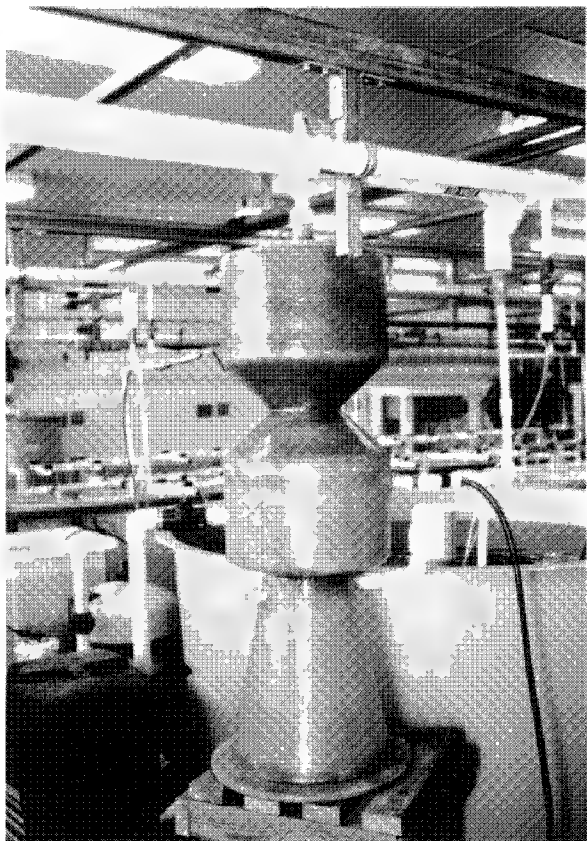


Fig. 7. A bead filter, one of many types of biological filters, shown in association with a laboratory-scale recirculating water system. Small plastic beads inside the fiber glass chamber provide surface area for colonization by bacteria that convert ammonia to nitrate.

Other than the biofilter and culture chambers, recirculating systems typically also employ one or more settling chambers or mechanical filters to remove solids such as unconsumed feed, feces, and mats of bacteria that slough from the biofilter into the water. Each recirculating system requires a mechanical means of moving water from component to component. That usually means mechanical pumping, though air-lifts can also be used.

Control of circulating bacteria and oxidation of organic matter can be obtained through ozonation of the water. Ozone (O_3) is highly toxic to aquatic organisms. Ozone must be allowed to dissipate prior to exposing the water to the aquaculture animals. With time, and with the assistance of aeration, ozone can be driven off or converted to molecular oxygen. Various commercial firms market ozone generators and can assist aquaculturists in selecting the proper equipment to meet system needs.

Ultraviolet (uv) light has also been used to sterilize the water in aquaculture systems. The effectiveness of uv decreases with the thickness of the water column being treated, so the water is usually flowed past uv lights as a thin film (alternatively, the water may flow through a tube a few cm in diameter that is surrounded by uv lights). Uv systems require more routine maintenance than ozone systems. Uv bulbs lose their power with time and need

to be changed periodically. In addition, organic material exposed to uv light foul the surface of the transparent quartz (sometimes plastic) tubes past or through which the water flows, thereby causing a film between the water and the uv source that reduces the effectiveness of the uv light.

Recirculating systems often feature other types of apparatus, such as foam strippers and supplemental aeration. The technology for denitrifying nitrate to nitrogen gas has developed to the point that it may find a place in commercial culture systems in the near future. Computerized water-quality monitoring systems that will sound alarms and call emergency telephone numbers to report systems failures to the culturists are finding increased use among culturists using recirculating systems.

The technology involved makes recirculating systems expensive to construct and operate. Redundancy in the system, ie, providing backups for all critical components, and automation are important considerations. When a pump fails, for example, the failure must be instantly communicated to the culturist and the culturist must have the ability to keep the system operating while the problem is being addressed. Loss of a critical component for even a few minutes can result in the loss of all animals within the system.

Recirculating systems can make aquaculture feasible in locations where conditions would otherwise not be conducive to successful operations. Such systems can also be used to reduce transportation costs by making it possible to grow animals near markets. In areas where there are concerns about pollution or the use of exotic species, closed systems provide an alternative approach to more extensive types of operations.

Water Sources and Quality

Sources of water for aquaculture include municipal supplies, wells, springs, streams, lakes, reservoirs, estuaries, and the ocean. The water may be used directly from the source or it may be treated in some fashion prior to use (see WATER).

Many municipal water sources are chlorinated and contain sufficiently high levels of chlorine so as to be toxic to aquatic life. Chlorine can be removed by passing the water through activated charcoal filters or through the use of sodium thiosulfate metered into the incoming water. Municipal water is usually not used in aquaculture operations that utilize large quantities of water, either continuously or periodically, because of the initial high cost of the water and the cost of pretreatment to remove chlorine.

If polled, most aquaculturists would probably indicate a preference for well water. Both freshwater and saline wells are common sources of water for aquaculture. The most commonly used pretreatments of well water include temperature alteration (either heating or cooling); aeration to add oxygen or to remove or oxidize such substances as carbon dioxide, hydrogen sulfide, and iron and increasing salinity (in mariculture systems). Pretreatment may also include adjusting pH, hardness, and alkalinity through the application of appropriate chemicals.

To heat or cool water requires large amounts of energy. A major consideration in locating an aquaculture facility is to have not only a sufficient supply of water, but to have water at or near the optimum temperature for growing the species that has been selected. The vast supply of spring water of almost perfect temperature in the Hagerman Valley of Idaho supports the majority of the trout production in the United States. Where geothermal water is available, tropical species can be grown in locations where ambient winter temperatures would otherwise not allow them to survive.

Another large cost associated with incoming water is associated with its movement. Many aquaculture facilities that utilize surface waters and those that obtain their water from wells other than artesian wells are required to pump the water into their facilities. Pumping costs can be a major expense, particularly when the facility requires continuous inflow.

Surface water can sometimes be obtained through gravity flow by locating aquaculture facilities at elevations below those of adjacent springs, streams, lakes, or reservoirs. Coastal facilities may be able to obtain water through tidal flow.

The most common treatment of incoming surface water is removal of particulate matter. This can be effected through the use of settling basins or filtration. Particle removal may involve the reduction or elimination of suspended inorganic material such as clay, silt, and sand. It may also involve removal of organic material, including living organisms. Organisms that enter aquaculture facilities if not filtered from the incoming water include phytoplankton and zooplankton, plants and plant parts, macroinvertebrates, and fishes. Some of the organisms, if not removed, can survive and grow to become predators on, or competitors with, the target aquaculture species. Very small organisms, such as bacteria, can be removed mechanically. However, other forms of sterilization, such as ozonation and the use of uv radiation, are more efficient and effective.

For many freshwater species that can be characterized as warmwater (such as channel catfish and tilapia) or coldwater (such as trout), the conditions outlined in Table 4 should provide an acceptable environment. So-called midrange species are those with an optimum temperature for growth of about 25°C (examples are walleye, northern pike, muskellunge, and yellow perch). Typically they do well under the conditions, other than temperature, specified in Table 4 for coldwater species. Some species have higher or lower tolerances than others. For example, tilapia can tolerate temperatures in excess of 34°C, but have poor tolerance for low temperature. Most tilapia species die when the temperature falls below about 12°C. Tilapia have a remarkably high tolerance for ammonia compared with such species as trout and salmon, which have a high tolerance for cold water, but cannot tolerate water temperatures much above 20°C. Marine fish may be able to tolerate a wide range of salinity (such euryhaline species include flounders, red drum, salmon, and some species of shrimp), or they may have a narrow tolerance range (they are called stenohaline species, examples of which are dolphin, halibut, and lobsters). Recommended water quality conditions for marine fish production systems are presented in Table 3.

Table 4. General Water Quality Requirements for Trout and Warmwater Aquatic Animals in Fresh Water^a

Variable	Acceptable level or range	
	Cold water	Warm water
temperature, °C	<20	26–30
alkalinity, mg L	10–400	50–400
dissolved oxygen, mg L	>5	>5
hardness, mg L	10–400	50–400
pH	6.5–8.5	6.5–8.5
total ammonia, mg L	<0.1	<1.0
ferrous iron, mg L	0	0
ferric iron, mg L	0.5	0–0.5
carbon dioxide, mg L	0–10	0–15
hydrogen sulfide, mg L	0	0
cadmium, µg L	<10	<10
chromium, µg L	<100	<100
copper, µg L	<25	<25
lead, µg L	<100	<100
mercury, µg L	<0.1	<0.1
zinc, µg L	<100	<100

^a Refs. 1, 6, 7.

The water quality criteria for each species should be determined from the literature or through experimentation when literature information is unavailable. Synergistic effects that occur among water quality variables can have an influence on the tolerance a species has under any given set of circumstances. Ammonia is a good example. Ionized ammonia (NH₄⁺) is not particularly lethal to aquatic animals, but unionized ammonia (NH₃) can be

toxic even when present at a fraction of a part per million (depending on species). The percentage of unionized ammonia in the water at any given total ammonia concentration changes in relation to such factors as temperature and pH. As either temperature or pH increase, so does the percentage of unionized ammonia relative to the level of total ammonia.

Another example is dissolved oxygen (DO). The amount of DO water can hold at saturation is affected by both temperature and salinity. The warmer and or saline the water, the lower the saturation DO level. Oxygen saturation is also affected by atmospheric pressure. The saturation oxygen level decreases as elevation increases.

Biocides should not be present in water used for aquaculture. Sources of herbicides and pesticides are runoff from agricultural land, contamination of the water table, and spray drift from crop-dusting activity. Excessive levels of phosphorus and nitrogen may occur where runoff from fertilized land enters an aquaculture facility either from surface runoff or groundwater contamination. Trace metal levels should be low as indicated in Tables 4 and 5.

Table 5. Suggested Water Quality Conditions for Marine Fish Production Facilities^a

Variable	Acceptable level or range
temperature, °C	1–40 (depends on species)
salinity, g kg	1–40 (depends on species)
dissolved oxygen, mg L	>6
pH	<7.9–8.2
total ammonia, µg L as NH ₃	<10
iron, µg L	100
carbon dioxide, mg L	<10
hydrogen sulfide, µg L	<1
cadmium, µg L	<3
chromium, µg L	<25
copper, µg L	<3
mercury, µg L	<0.1
nickel, µg L	<5
lead, µg L	<4
zinc, µg L	<25

^a Ref. 8.

Most aquaculture facilities release water constantly or periodically into the environment without passing it through a municipal sewage treatment plant. The effects of those effluents on natural systems have become a subject of intense scrutiny in recent years and have, in some instances, resulted in opposition to further development of aquaculture facilities in some locales. There have even been demands that some existing operations should be shut down.

Regulation of aquaculture varies greatly both between and within nations. Some governmental agencies with jurisdiction over aquaculture have placed severe restrictions on the levels of such nutrients as phosphorus and nitrogen that can be released into receiving waters. Regulations on suspended solids levels in effluent water are also common. The installation of settling ponds, exposure of the water to filter feeding animals that will remove solids, and mechanical filtration have been used to treat effluents. Reduction or removal of dissolved nutrients through tertiary treatment is possible, but is generally not economically feasible at present. Research is currently underway to develop feeds containing reduced levels of nutrients or to provide nutrients in forms that can better be utilized by the culture animals. The goal in both approaches is to reduce losses of nutrients to the environment through excretion.

Nutrition and Feeding

Problems associated with excessive levels of nutrients and unwanted nuisance species have already been mentioned. There are cases in which intentional fertilization is used by aquaculturists in order to produce desirable types of natural food for the species under culture. Examples of this approach include inorganic fertilizer applications in ponds to promote phytoplankton and zooplankton blooms that provide food for young fish such as channel catfish, the development of algal mats through fertilization of milkfish ponds, and the use of organic fertilizers (from livestock and human excrement) in Chinese carp ponds to encourage the growth of phytoplankton, macrophytes, and benthic invertebrates. In the latter instance, various species of carp with different food habits are stocked to ensure that all of the types of natural foods produced as a result of fertilization are consumed.

Provision of live foods is currently necessary for the early stages of many aquaculture species because acceptable prepared feeds have yet to be developed. Algae is routinely cultured for the early stages of molluscs produced in hatcheries. Once the molluscs are placed in growout areas, natural productivity is depended upon to provide the algae upon which the shellfish feed.

In cases where zooplankton are reared as a food for predatory larvae or fry, it may be necessary to maintain three cultures. Though wild zooplankton have been used successfully in some instances (eg, in Norway wild zooplankton have been collected and fed to larval Pacific halibut), the normal process involves culturing algae to feed to zooplankton that are fed to a young shrimp or fish.

The most popular live foods for first feeding animals such as shrimp and marine fishes that have small eggs and larvae are rotifers and brine shrimp nauplii. After periods ranging from several days to several weeks, depending on the species being reared, the aquaculture animals will become sufficiently large to accept pelleted feeds and can be weaned onto prepared diets. Problems associated with utilizing prepared feeds from first feeding include difficulty in providing very small particles that contain all the required nutrients, loss of soluble nutrients into the water from small particles before the animals consume the feed, and in some cases, the fact that prepared feeds do not behave the same as live foods when placed in the water. For species that are sight feeders, behavior of the food is an important factor.

Some of the most popular aquaculture species accept prepared feeds from first feeding. Included are catfish, tilapia, salmon, and trout. All of the fishes listed have relatively large eggs (several mm diameter) that develop into fry that have large yolk sacs. The nutrients in the yolk sac lead to production of first-feeding fry with well-developed digestive tracts that produce the enzymes required to efficiently digest diets that contain the same types of ingredients used for larger animals.

Fish nutritionists have, over the last few decades, successfully determined the nutritional requirements of many aquaculture species and have developed practical feed formulations based on those requirements. For species such as Atlantic salmon, various species of Pacific salmon and trout, common carp, channel catfish and tilapia sufficient information exists to design diets precisely suited to each species. There is always interest among aquaculturists to develop new species. In each instance, the nutritional requirements of the new species must be investigated. Although there are many similarities among aquatic animals, diets that produce the best growth at the least cost vary significantly and can only be formulated when the nutritional requirements are known. Determination of those requirements may require several years of research, although suitable diets based on existing formulations can be employed while the research is being conducted.

Requirements for energy, protein, carbohydrates, lipids, vitamins and minerals have been determined for the species commonly cultured (9). As a rule of thumb, trout and salmon diets will, if consumed, support growth and survival in virtually any aquaculture species. Such diets often serve as the control against which experimental diets are compared.

Since feeds contain other substances than those required by the animals of interest, studies have also been conducted on antinutritional factors in feedstuffs and on the use of additives. Certain feed ingredients contain chemicals that retard growth or may actually be toxic. Examples are gossypol in cottonseed meal and trypsin inhibitor in soybean meal. Restriction on the amount of the feedstuffs used is one way to avoid problems. In some cases, as is true of trypsin inhibitor, proper processing can destroy the antinutritional factor. In this case, heating of soybean meal is effective.

Animals that do not readily accept pelleted feeds may be enticed to do so if the feed carries an odor that induces ingestion. Color development is an important consideration in aquarium species and some animals produced for human food. External coloration is desired in aquarium species. Pink flesh in cultured salmon is desired by much of the consuming public. Coloration, whether external or of the flesh, can be achieved by incorporating ingredients that contain pigments or by adding extracts or synthetic compounds. One class of additives that imparts color is the carotenoids.

Prepared feeds are marketed in various forms from very fine particles through crumbles, flakes, and pellets. Pelleted rations may be hard, semimoist, or moist. Hard pellets typically contain less than 10% water and can be stored under cool, dry conditions for at least 90 days without deterioration of quality. Semimoist pellets are chemically stabilized to protect them from degradation and mold if they are properly stored, while moist pellets must be frozen if they are not used immediately after manufacture. Moist feeds are produced in machines similar to sausage grinders.

Hard pellets are the type preferred if the species under culture will accept them. Semimoist feeds are most commonly used in conjunction with feeding young fishes and species that find hard pellets unpalatable. Moist feeds, which contain high percentages of fresh fish, are usually available only in the vicinity of fish-processing plants.

The most widely used types of prepared feeds are produced by pressure pelleting or extrusion. Pressure pelleting involves pushing the ground and mixed feed ingredients through holes in a die that is a few centimeters thick to produce spaghetti-like strands of the desired diameter. The strands are cut to length as they exit the die. Steam is often injected into the pellet mill in a location that exposes the feed mixture to moist heat just before the mix enters the die. Exposure to steam improves binding and extends pellet water stability.

Extruded pellets are produced by exposing the ground and mixed ingredients to much higher heat and pressure and for a longer time than is the case with pressure pellets. In the extrusion process the ingredients undergo some cooking that can be beneficial in reducing the levels of certain antinutritional factors, such as trypsin inhibitor. There may be concomitant losses of heat labile nutrients such as vitamin C, so overfortification to obtain the desired level in the final product may be required.

Crumbles are formed by grinding pellets to the desired sizes. Specialty feeds such as flakes can be made by running newly manufactured pellets through a press or through use of a double drum dryer. The latter type of flakes begin as a slurry of feed ingredients and water. When the slurry is pressed between the hot rollers of the double drum dryer, wafer thin sheets of dry feed are produced that are then broken into small pieces. The different colors observed in some tropical fish foods represent a mixture of flakes, each of which contains one or more different additives that impart color.

Pressure pellets sink when placed in water, whereas under the proper conditions, floating pellets can be produced through the extrusion process. That is accomplished when the feed mixture contains high levels of starch that expands and traps air as the cooked pellets leave the barrel of the extruder. This gives the pellets a density of less than 1.0. Floating pellets are desirable for species that come to the surface to feed since the aquaculturist can visually determine that the fish are actively feeding and can control daily feeding rates based on observed consumption.

Sinking extruded pellets are used for shrimp and other species that will not surface to obtain food. Shrimp consume very small particles, so they will nibble pieces from a pellet over an extended period of time. For that reason, both pressure and extruded pellets need to have high water stability. Extruded feeds, whether sinking or floating, may remain intact for up to 24 hours after being placed in the water. Pressure pellets begin to disintegrate after a few minutes, unless supplemental binders are incorporated into the feed mixture. As previously indicated, the use of steam in conjunction with pressure pelleting also enhances pellet stability.

Nearly all aquaculture feeds contain at least some animal protein since the amino acid levels in plant proteins cannot meet the requirements of most aquatic animals. Fish meal is the most commonly used source of animal protein in aquaculture feeds, though blood meal, poultry by-product meal, and meat and bone meal have also been successfully used. Commonly used plant proteins include corn meal, cottonseed meal, peanut meal, rice, soybean meal, and wheat. A number of other ingredients have also been used, many of which are only locally available. Most formulations contain a few percent of added fat from such sources as fish oil, tallow, or more commonly, oilseed oils such as corn oil and soybean oil. Complete rations contain added vitamins and minerals. Purified amino acids, binders, carotenoids, and antioxidants are other components found in many feeds. Growth hormone and antibiotics are sometimes used. Regulations on the incorporation of hormones along with other chemicals and drugs into aquatic animal feeds are in place in the United States and some other countries (Table 6). Few such regulations have been promulgated in developing nations.

Table 6. Therapeutants and Disinfecting Agents Approved for Use in United States Aquaculture^a

Name of compound	Use of compound
<i>Therapeutants</i>	
copper	antibacterial for shrimp
formalin	parasiticide for various species
furanace (Nifurpyrinol)	antibiotic for aquarium fishes
oxytetracycline (Terramycin)	antibiotic for fishes and lobsters
sodium chloride	osmoregulatory enhancer for fishes
sulfadimethoxine (Romet)	antibacterial for salmonids and catfish
trichlorofon (Masoten)	parasiticide for baitfish and goldfish
<i>Disinfectants</i>	
calcium hypochlorite (HTH)	used in raceways and on equipment
didecyl dimethyl ammonium chloride (Sanaqua)	used in aquaria and fish-holding equipment
povidone–iodine compounds (Argentyne, Betadine, Wescodyne)	disinfection of fish eggs

^a Ref. 11.

Feeding practices vary from species to species. It is important not to overfeed since waste feed not only means wasted money, it can also lead to degradation of water quality. Most species require only three to four percent of body weight in dry feed daily for optimum growth. Very young animals are an exception. They are fed at a higher rate because they are growing rapidly and consume a greater percentage of body weight daily than older animals. It is important to have food readily available to them. Food should be spread evenly over the culture chamber area so the young animals do not have to expend a great deal of energy searching for a meal. Feeding rates as high as 50% of body weight daily are not uncommon for young animals. Since total biomass is small, even in intensively stocked units such as raceways, the economic cost is not high. Water quality in raceways can be maintained by siphoning out waste feed periodically. In ponds, any unconsumed feed acts as fertilizer and the quantities used are not high enough to affect water quality adversely.

Young animals may be fed several times daily. Examples include the standard practices of feeding fry channel catfish every three hours and young northern pike as frequently as every few minutes. Keeping carnivorous species such as northern pike satiated helps reduce the incidence of cannibalism.

Reproduction and Genetics

Species such as carp, salmon, trout, channel catfish, and tilapia have been bred for many generations in captivity though they usually differ little in appearance or genetically from their wild counterparts. A few exceptions exist, such as the leather carp, a common carp strain selectively bred to produce only one row of scales, and the Donaldson trout, a strain of rainbow trout developed over numerous generations to grow more rapidly to larger size and

with a stouter body than its wild cousins.

Selective breeding has been long-practiced as a mean of improving aquaculture stocks. In some instances it has not been possible or it is quite difficult and expensive to produce broodstock and spawn them in captivity, so culturists continue to rear animals obtained from nature. Most of the species that are being reared in significant quantities around the world are produced in hatcheries using either captured or cultured broodstock. Milkfish is a notable exception. The species has been spawned in captivity, but most of the fish reared in confinement are collected as juveniles in seines and sold to fish culturists. Wild shrimp post-larvae continue to be used to stock ponds in some parts of the world though hatcheries may also be available in the event sufficient numbers of wild post-larvae are unavailable in a given year.

Spawning techniques vary widely from one species to another. Tilapia and catfish are typically allowed to spawn in ponds. Fertilized eggs can be collected from the mouths of female tilapia, but it is common practice to collect schools of fry after they are released from the mother's mouth to forage on their own. Catfish lay eggs in adhesive masses. Spawning chambers such as milk cans and grease cans are placed in ponds and may be examined every few days for the presence of egg masses. Some catfish farmers allow the eggs to hatch in the pond, though most collect eggs and incubate them in a hatchery. Adult Pacific salmon die after spawning. Females are usually sacrificed by cutting open the abdomen to release the eggs. Milt is obtained by squeezing the belly of males. Trout and Atlantic salmon can be reconditioned to spawn annually. Eggs are usually obtained from those species in the same fashion as from male Pacific salmon (1).

Unlike catfish, tilapia, trout and salmon, that produce several hundred to several thousand eggs per female, many marine species produce large numbers of very small eggs. Hundreds of thousands to millions of eggs are produced by such species as halibut, flounders, red drum, striped bass, and shrimp. Catfish, salmon, and trout spawn once a year, while tilapia and some marine species spawn repeatedly if the proper environmental conditions are maintained (1). Red drum, for example, spawn every few days for periods of several months when light and temperature and properly controlled (10).

Fish breeders have worked with varying degrees of success to improve growth and disease resistance in a number of species. As genetic engineering techniques are adapted to aquatic animals, dramatic and rapid changes in the genetic makeup of aquaculture species may be possible. However, since it is virtually impossible to prevent escapement of aquacultured animals into the natural environment, potential negative impacts of such organisms on wild populations cannot be ignored.

For some species, one sex may grow more rapidly than the other. A prime example is tilapia, which mature at an early age (often within six months of hatching). At maturity, submarketable females divert large amounts of food energy to egg production. Also, since they are mouth brooders (holding the eggs and fry within their mouths for about two weeks) and repeat spawners (spawning about once a month if the water temperature is suitable), the females grow very slowly once they mature. All-male, or predominantly male, populations of tilapia can be produced by feeding androgens to fry, which are undifferentiated sexually. Various forms of testosterone have been used effectively in sex reversing tilapia and other fishes.

In species such as flatfishes, females may grow more rapidly than males and ultimately reach much larger sizes. For them, producing all-female populations for grow out might be beneficial.

Diseases and Their Control

Aquatic animals are susceptible to a variety of diseases including those caused by viruses, bacteria, fungi, and parasites. A range of chemicals and vaccines has been developed for treating the known diseases, although some conditions have resisted all control attempts to date and severe restrictions on the use of therapeutants in some nations has impaired that ability of aquaculturists to control disease outbreaks. The United States is a good example of a nation in which the variety of treatment chemicals is limited (Table 6).

Maintenance of conditions in the culture environment that keep stress to a minimum is one of the best methods of avoiding diseases. Vaccines have been developed against several diseases and more are under development. Selective breeding of animals with disease resistance has met with only limited success. Good sanitation and disinfection of contaminated facilities are important avoidance and control measure. Some disinfectants are listed in Table 6. Pond soils can be sterilized with burnt lime (CaO), hydrated lime [Ca(OH)₂], or chlorine compounds (12).

When treatment chemicals have to be employed, they may be incorporated in the food, used in dips, flushes and baths, or allowed to remain in the water for extended periods. Since one of the first responses of aquatic animals to disease is reduction or cessation of feeding, treatments with medicated feeds must be initiated as soon as development of an outbreak is suspected. Antibiotics, such as terramycin, can be dissolved in the water, but may be less effective than when given orally.

Vaccines can be administered through injections, orally, or by immersion. Injection is the most effective means of vaccinating aquatic animals but it is stressful, time-consuming, and expensive. The time and expense may be acceptable for use in conjunction with broodfish and other valuable animals. Oral administration of vaccines may be ineffective as many vaccines are deactivated in the digestive tract of the animals the vaccines are intended to protect. Dip treatment by which the vaccines enter the animals through diffusion from the water are not generally as effective as injection but can be used to vaccinate large numbers of animals in short periods of time.

Harvesting, Processing, and Marketing

Harvesting techniques vary depending on the type of culture system involved. Seines are often used to capture fish from ponds, or the majority of the animals can be collected by draining the pond through netting. Fish pumps are available that can physically remove aquatic animals directly onto hauling trucks from ponds, raceways, cages and net-pens without causing damage to the animals.

Aquaculturists may harvest, and even process their own crops, although custom harvesting and hauling companies are often available in areas where the aquaculture industry is sufficiently developed to support them. Some processing plants also provide harvesting and live-hauling services.

Some species, with channel catfish being a good example, can develop off-flavors. A characteristic off-flavor in catfish is often described as an earthy, musty, or muddy flavor. The problem is associated with the chemical geosmin and related compounds that are produced by certain types of algae (1). Processors often require that a sample fish from each pond scheduled for harvest be brought to the plant about two weeks prior to harvest for a taste test. Subsequent samples are taken to the processor three days before and during the day of processing. A portion of the sample fish is cooked and tasted. If off-flavor is detected, the fish will be rejected. Once the source of geosmin is no longer present, the fish will metabolize the compound. The process may involve moving the fish into clean well water or by merely waiting until the algae bloom dissipates, after which the geosmin will be rapidly metabolized. Within a few days the fish can be retested and if no off-flavor is detected, they can be harvested and processed.

Centralized processing plants specifically designed to handle regional aquaculture crops are established in areas where production is sufficiently high. In coastal regions, aquacultured animals are often processed in plants that also service capture fisheries.

Marketing can be done by aquaculturists who operate their own processing facilities. Most aquaculture operations depend on a regional processing plant to market the final product. In all cases aquaculturists should remember that their job is not complete until the product reaches the consumer in prime condition.

BIBLIOGRAPHY

"Aquaculture" in *ECT* 3rd ed., Vol. 3, pp. 194–213, by Howard P. Clemens and Michael Conway, University of Oklahoma.

1. R. R. Stickney, *Principles of Aquaculture*, John Wiley & Sons, Inc., New York, 1994.
2. FAO, *Agriculture Production Statistics*, 1974–1993, Food and Agriculture Organization of the United Nations, 1995.
3. FAO, *FAO Fisheries Circular 815, revision 2*, Food and Agriculture Organization of the United Nations, 1991.
4. *Aquaculture Buyer's Guide '95 and Industry Directory*, Vol. 8, 1995 Aquaculture Magazine, Asheville, NC.
5. R. R. Stickney, *Aquaculture in the United States: A Historical Review*, John Wiley & Sons, Inc., New York, 1996.

6.

R. G. Piper, I. B. McElwain, L. E. Orme, J. P. McCraren, L. G. Fowler, and J. R. Leonard, *Fish Hatchery Management*, U.S. Fish and Wildlife Service, Washington, D.C., 1982.

7.

C. E. Boyd, *Water Quality in Ponds for Aquaculture*, Alabama Agricultural Experiment State, Auburn University, 1990.

8.

J. Huegenin and J. Colt, *Design and Operating Guide for Aquaculture Seawater Systems*, Elsevier, New York, 1989.

9.

National Research Council, *Nutrient Requirements of Fish*, National Academy Press, Washington, D.C., 1993.

10.

A. Henderson-Arzapalo, *Rev. Aquat. Sci.* **6**, 479 (1992).

11.

F. P. Meyer and R. A. Schnick, *Rev. Aquatic Sci.* **1**, 693 (1989); a review of chemicals used for the control of fish diseases.

12.

C. E. Boyd, *Bottom Soils, Sediment, and Pond Aquaculture*, Chapman and Hall, New York, 1995.

Rorbert Stickney
Texas A&M University

BIOREMEDIATION

Bioremediation is the process of judiciously exploiting biological processes to minimize an unwanted environmental impact; usually it is the removal of a contaminant from the biosphere. Like most definitions in biology, that of bioremediation is the subject of some debate. A narrow definition might focus on the conversion of contaminating organic molecules to carbon dioxide, water, and inorganic ions, and the oxidation or reduction of contaminating inorganic ions. A broader definition would include biological processes for ameliorating extremes of pH, concentrating contaminants so that they can be more easily removed by physical techniques, converting toxic species to less toxic or less bioavailable forms that pose less of a threat to the environment, and restoring functional ecosystems to contaminated or disturbed sites when the contaminants or disturbance cannot be removed.

The concept of "judiciously exploiting biological approaches" is also a subject of debate:

Some would restrict it to providing a nutrient that is otherwise limiting the most effective growth of organisms catalyzing the desired reaction, whether it is fertilizer providing nitrogen, phosphorus and other essential minerals, or an electron acceptor such as oxygen.

Others would extend the fertilizer concept to the simultaneous addition of readily biodegradable substrates along with the fertilizer nutrients to stimulate the growth of contaminant-degrading organisms most rapidly, and to aid in the rapid utilization of the fertilizer nutrients before they might be leached from the contaminated area. The specific requirements for the most efficacious substrates is an area of current research.

An alternative use of added readily degradable substrates is to drive the local environment toward anaerobiosis so that reactions such as reductive dechlorinations or reductive removal of nitro-groups are promoted.

A broader view would include the addition of a substrate to stimulate the growth of organisms known to degrade the contaminant of interest only as an incidental part of their metabolism, one might almost say serendipitously. This process is sometimes called co-metabolism, and it too has had success.

Others would include the addition of materials aimed at increasing the bioavailability of the contaminant to the degrading organisms. The most studied compounds are surfactants, but cations have been reported to increase the bioavailability of some organic compounds, and sorbents and clays are also considered. The dispersion of spilled oil on water by the application of dispersants is perhaps the major commercial use of this idea.

Another important option is the addition of remediating organisms. While the addition of contaminant-degrading bacteria has not yet had much documented success with hydrocarbons, there is reason to expect that it will be efficacious with pollutants that are more recent additions to the environment. The planting of specific plants, in the process known as phytoremediation, is also a promising approach. Not only do the plants themselves have remediating activities, such as the accumulation of certain metal ions, but they also have extensive bacterial and fungal populations associated with their root systems, and inoculation of this rhizosphere is widely practiced. It is, thus, possible that planting seeds with microbial inoculants will become an option for bioremediation. There is much talk of genetically modifying organisms so that their remediative potential is increased, and in the future this may well become an important option.

The broadest view of "aiding and abetting" includes doing nothing, but merely watching natural processes occur without further intervention. This has been termed "Intrinsic Bioremediation", and it too has met with success. From an environmental point of view, although unfortunately not always from a regulatory viewpoint, it is important that any remediation intervention yield a clear net environmental benefit. Sometimes very mild stimulation of intrinsic processes may be the most environmentally responsible option.

Bioremediation overlaps some older biotechnologies. Municipal and industrial wastewater treatment is a well-established industry, and although it can be distinguished from bioremediation in that the pollutants are under physical control during treatment, the fundamental biological processes have much in common. Similarly composting is a well-established phenomenon, currently gaining popularity in the municipal solid waste treatment industry, and the biofiltration of waste gases is becoming a useful technology. Developments of these technologies, where the contaminant is already under physical control, will undoubtedly aid the development of bioremediation as an accepted tool for dealing with similar wastes when they have escaped control. This article focuses on biological treatments for contaminants when they have escaped into the environment.

Bioremediation is already a commercially viable technology, with estimates of aggregate bioremediation revenues of \$2–3 billion for the period 1994–2000 (1). There are significant opportunities to enlarge upon this success. Bioremediation has applications in the gas phase, in water, and in soils and sediments. For water and soils, the process can be carried out *in situ*, or after the contaminated medium has been moved to some sort of contained reactor (*ex situ*). The former is generally rather cheaper, but the latter may result in such a significant increase in rate that the additional cost of manipulating the contaminated material is overshadowed by the time saved. Bioremediation may explicitly exploit bacteria, fungi, algae, or higher plants. Each, in turn, may be part of a complex food-web, and optimizing the local ecosystem may be as important as focusing solely on the primary degraders or accumulators.

Bioremediation usually competes with alternative approaches to achieving an environmental goal. Bioremediation is typically among the least expensive options, but an additional important consideration is that in many cases bioremediation is a permanent solution to the contamination problem, since the contaminant is completely destroyed or collected. Some of the alternative technologies, such as thermal desorption and destruction of organics, are also permanent solutions, but the simplest, removing the contaminant to a dump site, merely moves the problem, and may well not eliminate the potential liability. Furthermore, by its very nature bioremediation addresses the bioavailable part of any contamination, and when biodegradation or bioaccumulation ceases this probably means that the bioavailable part of the contamination has been addressed. Residual concentrations of contaminants, although perhaps detectable by today's sensitive analytical techniques, may in fact have no residual environmental impact. The same cannot necessarily be said for nonbiological technologies, which may leave bioavailable contaminants at low levels.

Bioremediation also has the advantage that it can be relatively nonintrusive, and can sometimes be used in situations where other approaches would be severely disruptive. For example, bioremediation has been used to clean up hydrocarbon spills under buildings, roads, and airport runways without interfering with the continued use of these facilities.

On the other hand, bioremediation is usually slower than most physical techniques, and may not always be able to meet some very strict clean-up standards. Nevertheless, it is becoming a widely used technology. This article addresses bioremediation in its broadest sense, focusing on the contaminants that can be treated, the underlying biological processes that can mitigate the contamination, and the technologies that have been used, or are being developed, to treat them.

General Biological Aspects

The biosphere plays an important role in the great elemental cycles of the earth (2), and bioremediation must be placed in this context if it is to be appreciated in its broadest ramifications. One of the underlying fundamental truths of biological diversity is that if there is free energy available in the metabolism of a substrate, there is probably a guild of organisms that has evolved to make use of it. This is particularly germane to the biodegradation of organic molecules. For example, crude oil seeps to both land and water have occurred for millennia, and as a consequence, aerobic oil-degrading microorganisms are ubiquitous. If biology does not yet take advantage of a source of free energy, then it can be expected that there will be a strong selection pressure in favor of any organism that develops an ability to exploit it. This has been seen with by-products of nylon manufacture, where a *Pseudomonas aeruginosa* has gained the ability to degrade the novel compound 6-aminohexanoate linear dimer, a by-product of nylon-6 manufacture, as the

sole source of carbon and nitrogen (3). The successful bioremediation of xenobiotic compounds, such as pesticides and herbicides, may well represent a similar acquisition of traits.

Not all organic molecules provide a source of free energy, however. Some, such as small halogenated solvents, provide no significant source of nutrients or energy, and their aerobic destruction can only occur co-metabolically with the degradation of a more nutritious substrate (4). The white-rot fungi provide another variation on this theme. These organisms seem unique in their ability to degrade lignin, the structural polymer of higher plants. They may not gain any direct energetic benefit from lignin degradation, but it clearly allows access to cellulose which is a substrate for growth. Lignin degradation is catalyzed by a group of extracellular peroxidases that generate nonspecific oxidants, and there have been several proposals to use these systems for destroying contaminating organic compounds (5).

With successful bioremediation, organic compounds can eventually be converted to carbon dioxide, water, and biomass. Similarly, nitrogenous molecules, such as excess ammonia or nitrate in groundwater, can be mineralized to gaseous nitrogen. Alternatively they can stimulate the growth of plants, either terrestrial or marine, and the plant biomass can eventually be harvested so that the nitrogen is effectively removed from the local environment. Other nonorganic contaminants provide a different challenge for bioremediation. A few, such as mercury and selenium, are volatilized by some biological processes, but it is not clear that this is always beneficial. In some cases, such as chromium and arsenic, there is a dramatic difference in environmental toxicity depending on the redox state of the contaminant. Bioremediation has sometimes focused on this detoxification, usually by bacterial processes. A more satisfying approach would be to use a biological process to accumulate and concentrate the contaminant so that it can be removed for safe disposal. Fungi, algae, and higher plants have all been used in these efforts.

Table 1 explains a few of the biological terms that are widely used in discussing bioremediation, and which are used in the following text.

Table 1. Some Biological Definitions Relevant to Bioremediation

Term	Explanation
aerobic	conditions with free oxygen
anaerobic	conditions scrupulously free of oxygen
anoxic	conditions with very low levels of oxygen
autotrophic	growth using atmospheric CO ₂ as sole source of carbon
co-metabolic degradation	biodegradation of a contaminant only fortuitously with degradation of a true substrate
denitrification	the reduction of nitrate to gaseous nitrogen
Eukaryotes	organisms with a membrane-bound nucleus; the protozoa, fungi, plants, and animals
eutrophic	very rich nutrient conditions, especially of nitrogen compounds
heterotrophic	growth at the expense of complex organic substrates
lignolytic	growth of white-rot fungi under conditions where they synthesize lignin-degrading peroxidases
methanogenic	very anaerobic conditions, where carbon dioxide is reduced to methane
methanotrophic	aerobic growth with methane as sole source of carbon and energy
mineralization	conversion of a contaminant to its simplest forms, eg, CO ₂ , H ₂ O, CH ₄ , N ₂ , Cl
nitrate-reducing, denitrifying	anoxic conditions, where nitrate is reduced to nitrogen gas
nitrification	the biological oxidation of ammonia to nitrite and nitrate
oligotrophic	very low nutrient conditions
Prokaryotes	organisms lacking a membrane-bound nucleus; the bacteria and archaea
recalcitrant	very resistant to biodegradation
reductive dehalogenation, reductive dechlorination	the sequential loss of halogen substituents under anaerobic, usually methanogenic, conditions
rhizosphere	the soil around plant roots; this zone has different microbial populations from the bulk soil
sulfate-reducing	very anaerobic conditions, where sulfate is reduced, by sulfate-reducing bacteria, to sulfide
vadose zone	the part of the soil above the water table

General Technological Aspects

Successful bioremediation hinges upon the effective application of the biology discussed above. Sometimes the contaminant is on the surface, so access to it is reasonably simple. Indeed the required technology may be as simple as broadcast spreaders or sprayers to apply fertilizers, or tilling the soil to allow good aeration. Of course this is not necessarily as simple as it sounds, since contaminated sites are often very different from agricultural fields, and the technology has to be significantly stronger to "plow" the soil. Frequently the contaminant is below the surface, and applying even simple bioremediation strategies can be very involved. Table 2 lists some of the technologies in use today.

Table 2. Some Technological Definitions Relevant to Bioremediation

Technology	Description
air sparging; aquifer sparging; biosparging	injection of air to stimulate aerobic degradation; may also stimulate volatilization
air stripping	injection of air to stimulate volatilization
aquifer bioremediation	<i>in situ</i> bioremediation in an aquifer, usually by adding nutrients or co-substrates
aquifer sparging	injection of air into a contaminated aquifer to stimulate aerobic degradation, may also stimulate volatilization
batch reactor	a bioreactor loaded with contaminated material, and run until the contaminant has been consumed, then emptied, and the process is repeated
bioactive barrier; bioactive zone; biowall	a zone, usually subsurface, where biodegradation of a contaminant occurs so that no contaminant passes the barrier
bioaugmentation	addition of exogenous bacteria with defined degradation potential (or rarely indigenous bacteria cultivated in a reactor and reapplied)
biofilm reactor	a reactor where bacterial communities are encouraged on a high surface area support, biofilms often have a redox gradient so that the deepest layer is anaerobic while the outside is aerobic
biofiltration	usually an air filter with degrading organisms supported on a high surface area support such as granulated activated carbon
biofluffing	augering soil to increase porosity
bioleaching	extracting metallic contaminants at acid pH
biological fluidized bed; fluidized-bed bioreactor	bioreactor where the fluid phase is moving fast enough to suspend the solid phase as a fluid-like phase
biopile; soil heaping	an engineered pile of excavated contaminated soil, with engineering to optimize air,

bioslurping	water, and nutrient control vacuum extraction of the floating contaminant, water, and vapor from the vadose zone; the air flow stimulates biodegradation
biostimulation	optimizing conditions for the indigenous biota to degrade the contaminant
biotransformation	the biological conversion of a contaminant to some other form, but not to carbon dioxide and water
biotrickling filter	a reactor where a contaminated gas stream passes up a reactor with immobilized micro-organisms on a solid support, while nutrient liquor trickles down the reactor
bioventing	vacuum extraction of contaminant vapors from the vadose zone,thereby drawing in air that stimulates the biodegradation of the remainder
borehole bioreactor; in-well bioreactor	the addition of nutrients and electron acceptor to stimulate biodegradation <i>in situ</i> in a contaminated aquifer
closed-loop bioremediation	groundwater recovery, a bioreactor, and low-pressure reinjection to maximize nutrient use, and maintain temperature in cold climates
composting	addition of biodegradable bulking agent to stimulate microbial activity; optimal composting generally involves self-heating to 50–60°C
constructed wetland	artificial marsh for bioremediation of contaminated water
continuous stirred tank reactor (CSTR)	a completely mixed bioreactor
digester	usually an anaerobic bioreactor for digestion of solids and sludges that generates methane
<i>ex-situ</i> bioremediation	usually the bioremediation of excavated contaminated soil in a biopile, compost system or bioreactor
fixed-bed bioreactor	bioreactor with immobilized cells on a packed column matrix
land-farming; land treatment	application of a biodegradable sludge as a thin layer to a soil to encourage biodegradation; the soil is typically tilled regularly
natural attenuation; intrinsic bioremediation	unassisted biodegradation of a contaminant
phytoextraction	the use of plants to remove and accumulate contaminants from soil or water to harvestable biomass
phytofiltration	the use of completely immersed plant seedlings, to remove contaminants from water
phytoremediation	the use of plants to effect bioremediation
phytostabilization	the use of plants to stabilize soil against wind and water erosion
pump and treat	pumping groundwater to the surface, treating, and reinjection or disposal
rhizofiltration	the use of roots to immobilize contaminants from a water stream
rotating biological contactor	bioreactor with rotating device that moves a biofilm through the bulk water phase and the air phase to stimulate aerobic degradation
sequencing batch reactor	periodically aerated solid phase or slurry bioreactor operated in batch mode
soil-vapor extraction	vacuum-assisted vapor extraction

Organic Contaminants

Hydrocarbons.

Constituents. Hydrocarbons get into the environment from biogenic and fossil sources. Methane is produced by anaerobic bacteria in enormous quantities in soils, sediments, ruminants and termites, and it is consumed by methanotrophic bacteria on a similar scale. Submarine methane seeps support substantial oases of marine life, with a variety of invertebrates possessing symbiotic methanotrophic bacteria (6). Thus, methanotrophic bacteria are ubiquitous in aerobic environments. Plants generate large amounts of volatile hydrocarbons, including isoprene and a range of terpenes (7). These compounds provide an abundant substrate for hydrocarbon-degrading organisms.

Crude oil has been part of the biosphere for millennia, leaking from oil seeps on land and in the sea. Crude oils are very complex mixtures, primarily of hydrocarbons although some components do have heteroatoms such as nitrogen (eg, carbazole) or sulfur (eg, dibenzothiophene). Chemically, the principal components of crude oils and refined products can be classified as aliphatics, aromatics, naphthenics, and asphaltic molecules. Representative examples are shown in Figure 1. The ratios of these different classes varies in different oils, but a typical crude oil might contain the four classes in a ratio of approximately 30:30:30:10. Most crude oils contain hydrocarbons ranging in size from methane to molecules with hundreds of carbons, although the lightest molecules are usually absent in oils that have been partially biodegraded in their reservoir. When crude oils reach the surface environment the lighter molecules evaporate, and are either destroyed by atmospheric photooxidation or are washed out of the atmosphere in rain, and are biodegraded. Some molecules, such as the smaller aromatics (benzene, toluene, etc) have significant solubilities, and can be washed out of floating slicks, whether these are at sea, or on terrestrial water tables. Fortunately the majority of molecules in crude oils, and refined products made from them, are biodegradable, at least under aerobic conditions.

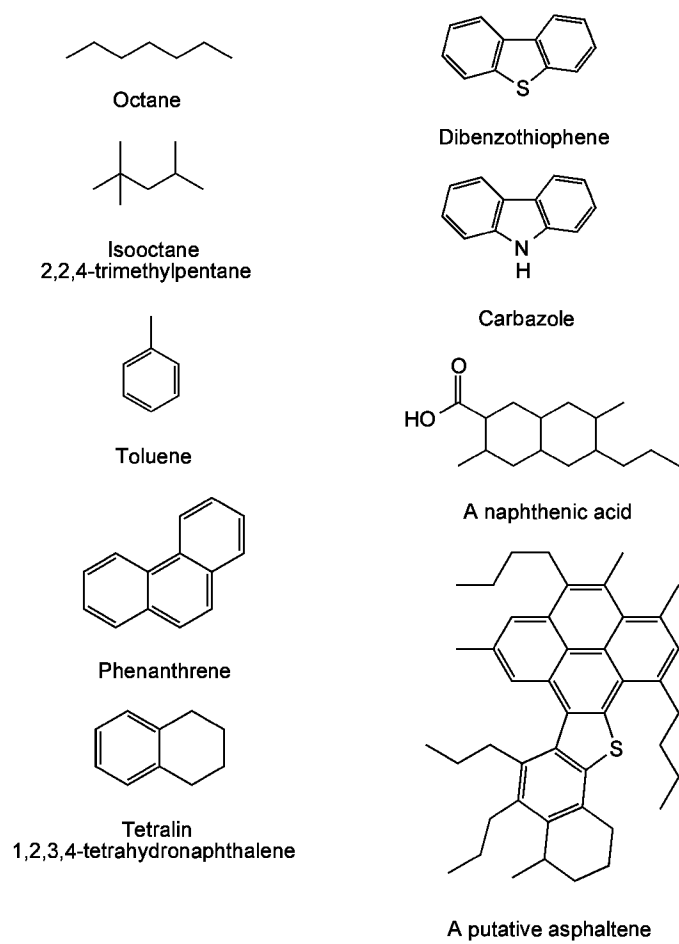


Fig. 1. Some representative hydrocarbons found in crude oil.

Biodegradation. Methane and the volatile plant terpenes are fully biodegradable by aerobic organisms, and most refined petroleum products are essentially completely biodegradable under aerobic conditions. Estimates for crude oil biodegradability range up to 90% (8), and the least biodegradable material, principally polar molecules and asphaltenes, lacks the "oily" feel and properties that are associated with oil. These are essentially impossible to distinguish from more recent organic material in soils and sediments, such as the humic and fulvic acids, and appear to be biologically inert.

Numerous bacterial and fungal genera have species able to degrade hydrocarbons aerobically and the pathways of degradation of representative aliphatic, naphthenic and aromatic molecules have been well characterized in at least some species (8). Other organisms, such as algae and plants, do not seem to play a very important role in the biodegradation of hydrocarbons. It is a truism that the hallmark of an oil-degrading organism is its ability to insert oxygen atoms into the hydrocarbon, and there are many ways in which this is achieved. Figures 2 and 3 show the most well-studied. Once a hydrocarbon possesses a carboxylate or alcohol functionality it is almost invariably a readily degradable compound. A simple example at the human level is the difference between oleic acid, a high calorie food, and octadecane, present in mineral oil, which is so inert that it serves as an intestinal lubricant!

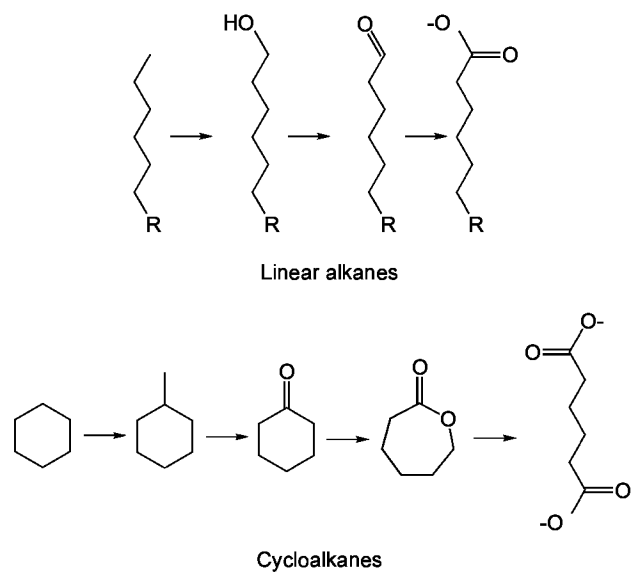


Fig. 2. Initial steps in the biodegradation of linear and cyclic alkanes.

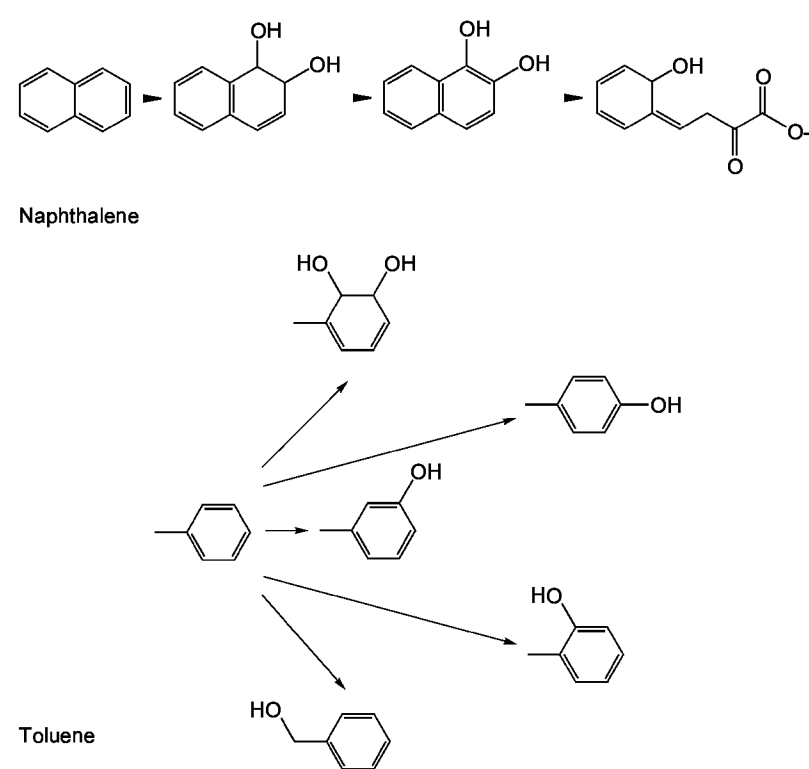


Fig. 3. Initial steps in the aerobic degradation of naphthalene, as a representative multiringed aromatic, and toluene. The different initial steps of toluene degradation are examples of the diversity found in different organisms.

For many years it was assumed that oil biodegradation was an exclusively aerobic process, since any degradation must involve oxidation. Indeed the very existence of oil reservoirs indicates that anaerobic degradative processes in such environments must be very slow. Nevertheless, in recent years it has become clear that at least some hydrocarbons are oxidized by bacteria under completely anaerobic conditions, where the oxygen is probably coming from water. Limited hydrocarbon biodegradation has now been shown under sulfate-, nitrate-, carbon dioxide- and ferric iron-reducing conditions (Table 3). The phenomenon is still poorly understood, however, and at present the largest molecules demonstrated to undergo biodegradation under these conditions are hexadecane, heptadecene, and phenanthrene. The pathways of degradation are only beginning to be addressed. Figure 4 shows the intermediates identified in anaerobic toluene degradation in different organisms. It is noteworthy that while organisms capable of aerobic oil biodegradation seem to be ubiquitous, organisms capable of the anaerobic degradation of hydrocarbon have to date only been found in a few places.

Table 3. Hydrocarbons That Have Been Shown to be Biodegraded Under Anaerobic Conditions

Electron acceptor	Substrate
nitrate (to nitrogen)	heptadecene toluene, ethylbenzene, xylene naphthalene
iron(III) (to iron(II))	terpenes
manganese(IV) (to Mn(II))	toluene
sulfate (to sulfide)	toluene hexadecane, alkylbenzenes benzene naphthalene, phenanthrene
CO ₂ (to methane)	toluene, xylene

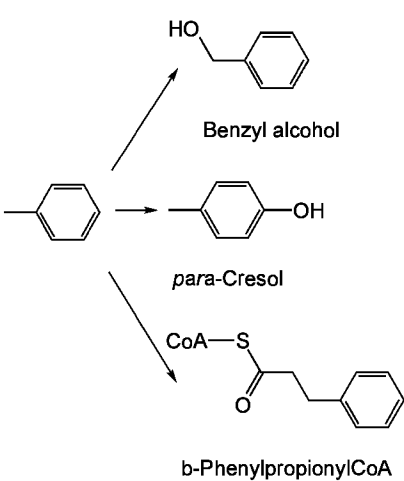


Fig. 4. Proposed initial steps in the anaerobic biodegradation of toluene in different organisms.

Although the majority of molecules in crude oils and refined products are hydrocarbons, the U.S. Clean Air Act amendment of 1990 mandated the addition of oxygenated compounds to gasoline in many parts of the United States. The requirement is usually that 2% (w/w) of the fuel be oxygen, which

requires that 5–15% (v/v) of the gasoline be an oxygenated additive (eg, methanol, ethanol, methyl *tert*-butyl ether (MTBE), etc). Although methanol and ethanol are readily degraded under aerobic conditions, the degradability of MTBE remains something of an open question. The compound was previously very rare in the environment, but now it is one of the major chemicals in commerce. At first it seemed that the compound was completely resistant to biodegradation, but complete mineralization has now been reported (9). Whether biodegradation can be optimized for effective bioremediation remains to be seen.

Bioremediation. Crude oil and refined products are readily biodegradable under aerobic conditions, but they are only incomplete foods since they lack any significant nitrogen, phosphorus, and essential trace elements. Bioremediation strategies for removing large quantities of hydrocarbon must therefore include the addition of fertilizers to provide these elements in a bioavailable form.

Air. Hydrocarbon vapors in air are readily treated with biofilters. These are typically rather large devices with a very large surface area provided by bulky material such as a bark or straw compost. The contaminated air, perhaps from a soil vapor-extraction treatment, or from a factory using hydrocarbon solvents, is blown through the filter, and organisms, usually indigenous to the filter material or provided by a soil or commercial inoculum, grow and consume the hydrocarbons. Adequate moisture must be maintained for effective operation. Alternatively, trickling biofilters with recycled water are also in use. Both bacteria and fungi readily colonize such filters, and they can be very effective. Nevertheless, biofilters are usually equipped with a small granulated activated carbon "backup" filter to handle any sudden pulse loads that might overwhelm the biological capacity of the filter. Biofilters compete with granulated activated carbon filters, and are often cheaper because they minimize the cost of the granulated activated carbon, and the energy required to destroy the contaminant and the granulated activated carbon when the latter is saturated (10). Potential problems include plugging and uneven air or water flow, but successful designs work for many years with minimal maintenance except the occasional addition of nutrients and stirring of the bed.

Sea. Crude oil spills at sea are perhaps the most widely covered environmental incidents in the national and international media. Despite their notoriety, catastrophic tanker spills and well blow-outs are fortunately rather rare, and their total input into the world's oceans is approximately equivalent to that from natural seeps; significantly more oil reaches the world's oceans from municipal sewers (11). Physical collection of the spilled oil is the preferred remediation option, but if skimming is unable to collect the oil, biodegradation and perhaps combustion or photooxidation are the only routes for elimination of the spill. One approach to stimulating biodegradation is to disperse the oil with chemical dispersants. Early dispersants had undesirable toxicity, but modern dispersants and application protocols can stimulate biodegradation by increasing the surface area of the oil available for microbial attachment, and perhaps providing nutrients to stimulate microbial growth (12). Patents have been issued for dispersant formulations that specifically include nitrogen and phosphorus nutrients (13), but the products are not currently commercially available.

Bioremediation by the addition of oil-degrading microbes is often promoted as a treatment option for floating spills, but this approach has not yet met with any documented success (13).

Shorelines. The successful bioremediation of shorelines affected by the spill from the *Exxon Valdez* in Prince William Sound, Alaska, was perhaps the largest bioremediation project to date (14,15). More than 73 miles of shoreline were treated in 1989 and similar amounts of fertilizer were used in 1990. Oil had typically penetrated into the surface gravel on these shorelines, occasionally getting as deep as 30 cm into the sediment. Since the gravel was typically very permeable, oxygen availability was unlikely to be the limiting factor for biodegradation, and indeed this was subsequently shown to be correct. Bioremediation thus focused on the addition of nitrogen and phosphorus fertilizers to partially remove the nutrient-limitation on oil degradation. Of course the addition of fertilizers was complicated by the fact that oiled shorelines were washed by tides twice a day. These tides would have rapidly removed any soluble fertilizer, so a strategy was sought that would provide nutrients for a significant length of time. Various approaches to applying fertilizers were tried, including both standard and slow release nutrients, oleophilic nutrients and solutions of liquid fertilizers. Two fertilizers were used in the full-scale applications; one, an oleophilic product known as Inipol EAP22 (trademark of CECA, Paris, France), was a microemulsion of a concentrated solution of urea in an oil phase of oleic acid and triaurethylphosphate, with butoxyethanol as a cosolvent. This product was designed to adhere to oil, and to release its nutrients to bacteria growing at the oil-water interface. The other fertilizer was a slow-release formulation of inorganic nutrients, primarily ammonium nitrate and ammonium phosphate, in a polymerized vegetable oil skin. This product, known as Customblen (trademark of Grace-Sierra, Milpitas, California), released nutrients with every tide, and these were distributed throughout the oiled zone as the tide fell. Fertilizer application rates were carefully monitored so that the nutrients would cause no harm, and the rate of oil biodegradation was stimulated between two- and five-fold (14,15).

A wide range of fertilizers, including agricultural and horticultural fertilizers, and bone and fish meals have been tried at the pilot scale, usually with at least modest success (13). Some current work is aimed at addressing whether providing a readily degradable substrate with the fertilizer nutrients helps immobilize the nutrients in biomass at the oiled site. Of course nutrient-supplementation is only likely to markedly stimulate the rate of biodegradation where nutrient levels are naturally low. It is unlikely that fertilizers will have a dramatic effect in situations where agricultural or municipal run-off maintains elevated levels of nutrients, such as happens in some estuaries and bays. Here aeration is likely to be most effective.

Bioremediation by the addition of oil-degrading microbes has been promoted as a treatment option for oiled shorelines, but this approach has not yet met with any documented success (13).

Areas where there are currently few remediation options include oiled marshes, mangroves, and coral reefs. These environments are generally easily damaged by human intrusion and physical cleaning options may not provide any net environmental benefit. Bioremediation may provide some attractive options, and some success has been claimed, on a small scale, with fertilizer applications (13). Marshes and mangroves offer the additional complication that they are typically anoxic. Perhaps the anaerobic degradation of oil could be stimulated by inoculation with anaerobic hydrocarbon degrading microbes, or perhaps gentle aeration or the addition of slow release oxygen compounds, such as some inorganic peroxides, might stimulate aerobic degradation without significantly changing the redox balance of these environments. This is an area where research is very much in its infancy and there are no well-documented success stories to date.

Bioremediation also offers options for dealing with oiled material, such as seaweed, that gets stranded on shorelines; composting has been shown to be effective.

Groundwater. Spills of refined petroleum product on land, and leaking underground storage tanks, sometimes contaminate groundwater. Bioremediation is becoming an increasingly popular treatment for such situations.

Hydrocarbons typically have a specific gravity of less than 1, and refined products usually float on the water table if they penetrate soil that deeply. In the parlance of the remediation industry, such floating spills are often called NAPLs (nonaqueous phase liquids). Indeed they are sometimes known as LNAPLs for light nonaqueous phase liquids, to distinguish them from more dense materials, such as halogenated compounds, which are more likely to sink in groundwater. Stand-alone bioremediation is an option for these situations, but "pump and treat" is the more usual treatment. Contaminated water is brought to the surface, free product is removed by flotation, and the cleaned water re-injected into the aquifer or discarded. Adding a bioremediation component to the treatment, typically by adding oxygen and low levels of nutrients, is an appealing and cost-effective way of stimulating the degradation of the residual hydrocarbon not extracted by the pumping. This approach is becoming widely used.

Hydrocarbons are not very soluble in water, but the most soluble components leach out of a spill if there is continual flushing. Typically only small aromatic molecules, the infamous BTEX (benzene, toluene, ethylbenzene, and xylenes, Fig. 5), are soluble enough to contaminate groundwater. Although with the advent of oxygenated gasolines, it is expected that these oxygenates (ethanol, methanol, MTBE (methyl-*tert*-butyl ether) etc) will also be found in groundwater. In the past, remediation of such situations has usually used pump-and-treat methodologies. These methods are slow and may leave reservoirs of contaminants in pockets that are poorly connected to the main water body. Of course the contaminant is biodegradable, and some biodegradation is probably already occurring when the contamination is discovered. The cheapest approach to remediation is, thus, to allow this intrinsic process to continue. Evidence that it is indeed occurring can be found in the selective disappearance of the most biodegradable compounds in the contaminant mixture, and the concomitant disappearance of electron acceptors from the groundwater. Thus oxygen is depleted as the preferential terminal electron acceptor for metabolism, followed by nitrate, ferric iron, sulfate, and finally CO₂ for methanogenesis (16).

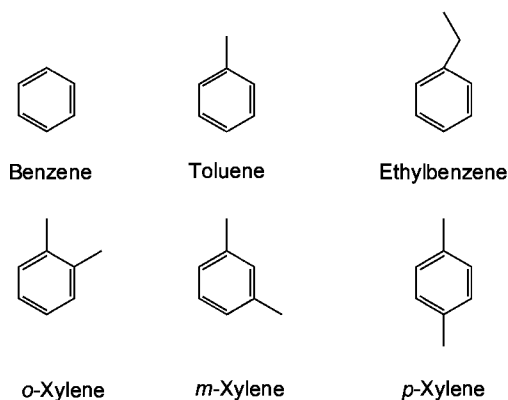


Fig. 5. The components of BTEX.

Intrinsic bioremediation is becoming an acceptable option in locations where the contaminated groundwater poses little threat to environmental health. Nevertheless, although intrinsic bioremediation is appealingly simple, it may not be the lowest cost option if there are extensive monitoring and documentation costs involved for several years. In such cases it may well be more cost effective to optimize conditions for biodegradation.

One approach is to optimize the levels of electron acceptors. Oxygen can be pumped in as the pure gas, or as air, although this is relatively energy intensive since oxygen is so poorly soluble. Hydrogen peroxide has been used in some situations, but there have been problems with biomass plugging near the injection wells. Slow release formulations of inorganic peroxides, such as magnesium peroxide, have recently been used with success (17). Nitrate may be added, although there are sometimes regulatory limitations on the amount of this material that may be added to groundwater (18). Ferric iron availability may be manipulated by adding ligands (19).

If there are significant amounts of both volatile and nonvolatile contaminants, remediation may be achieved by a combination of liquid and vapor extraction of the former, and bioremediation of the latter. This combination has been termed "bioslurping", where the act of pumping out the liquid contaminant phase draws in air at other wells to stimulate aerobic degradation (20). Such bioremediation requires that there be enough nutrients to allow microbial growth, and fertilizer nutrients are frequently added at the air injection wells. Bioslurping has had a number of well-documented successes.

The majority of remediation operations include stopping the source of the contamination, but in some cases this is impossible, either because of the location of the spill, or because it is over a large area, and not a point source. In these situations it may be possible to intercept the flow of contaminated groundwater off-site, and ensure that no contamination passes. The simplest intervention is a line of wells for pump and treat, but including a biological component may be more cost effective. This can range from the installation of a sparge line for aerating the contaminated plume, to installing some form of semicontained bioreactor where nutrients can be applied with some modicum of control. Often these designs are combined with barriers to ensure that all the contaminated plume passes through the reactive zone. These designs have a variety of names, including biowall, trench biosparge, funnel and gate, bubble curtain, sparge curtain and engineered trenches and gates. Both aerobic and anaerobic designs have been successfully installed.

Where there are large volumes of contaminated water under a small site, it is sometimes most convenient to treat the contaminant in a biological reactor at the surface. Considerable research has gone into reactor optimization for different situations and a variety of stirred reactors, fluidized-bed reactors, and trickling filters have been developed. Such reactors are usually much more efficient than *in situ* treatments, although correspondingly more expensive.

Of course the presence of a liquid phase of hydrocarbon in a soil gives rise to vapor contamination in the vadose zone above the water table. This can be treated by vacuum extraction, and the passage of the exhaust gases through a biofilter (see above) can be a cheap and effective way of destroying the contaminant permanently.

Soil. Hydrocarbon contamination of soils runs the gamut from crude oils at production well and pipeline spills, to the full slate of refined products at refineries, distribution centers, service stations and accident sites. Significant hydrocarbon contamination is also often found at manufactured gas plants, now mainly abandoned, wood treatment facilities, railroad rights of way and terminals, and various military bases. Sometimes the contamination is the result of leaking underground storage tanks and pipelines, leading to subsurface contamination, but surface spills also occur. Physical removal of gross contamination is an obvious first step at all locations, and bioremediation is an appealing option for remediating residual contamination in many of these sites.

Spills from production facilities and pipelines often involve both oil and brine, since most oil reservoirs float on top of concentrated brines, and both are produced in later stages of production. The brine is typically separated from the oil and re-injected into the reservoir, but some is retained in many production pipelines. The environmental impact of spilled brine can be quite deleterious. Not only is salt toxic to most plants, and can inhibit many soil bacteria, but it also can have a major effect on the soil structure by altering the physical properties of clays. Successful bioremediation strategies must therefore include remediating the brine. In wet regions the salt is eventually diluted by rainfall, but in arid regions, and to speed the process in wetter regions, gypsum is often added to restore soil porosity.

Many hydrocarbons bind quite tightly to soil components, and are thereby less available to microbial degradation. The kinetics of binding seem to be complex, and the process of "aging" is only poorly understood. Nevertheless, it seems clear that hydrocarbons that have been in contact with soil for a long time are not as available for biodegradation as fresh spills. Several groups of researchers have suggested the addition of surfactants to overcome this limitation, but this approach is not yet widely used. A significant potential concern is that the surfactant will be degraded in preference to the contaminant of concern.

Intrinsic biodegradation occurs, but it usually only removes the lightest refined products, such as gasoline, diesel and jet fuel. Active intervention is typically required. Usually the least expensive approach is *in situ* remediation, typically with the addition of nutrients, and the attempted optimization of moisture and oxygen by tilling. Various approaches to applying fertilizers have been tried, including both standard and slow-release nutrients, oleophilic nutrients and solutions of liquid fertilizers. Oxygen is a likely limiting nutrient in many cases, and soil tilling is widely practiced. This *in situ* bioremediation of hydrocarbon-contaminated soils is akin to the old practice of "land-farming", wherein sludges and other refinery wastes were deliberately spread onto soil and tilled and fertilized to stimulate biodegradation. Although this practice is now discontinued in the United States, it was quite widely used.

Deeper contamination may be remedied with bioventing, where air is injected through some wells, and extracted through others to both strip volatiles and provide oxygen to indigenous organisms. Fertilizer nutrients may also be added. This is usually only a viable option with lighter refined products.

A recent suggestion has been to use plants to stimulate the microbial degradation of the hydrocarbon (hydrocarbon phytoremediation). This has yet to receive clear experimental verification, but the plants are proposed to help deliver air to the soil microbes, and to stimulate microbial growth in the rhizosphere by the release of nutrients from the roots. The esthetic appeal of an active phytoremediation project can be very great.

When soil contamination extends to some depth it may be preferable to excavate the contaminated soil and put it into "biopiles" where oxygen, nutrient and moisture levels are more easily controlled. Biopiles can also be kept warm during winter months, increasing the amount of time available for biodegradation in colder climates. Since the soil is well mixed during the construction of the pile, there is an opportunity to add selected microbial and fungal strains in an additional attempt to maximize biodegradation.

Composting by the addition of readily degradable bulking agents is also a useful option for relatively small volumes of excavated contaminated soil.

Since efficient composting invariably involves self-heating as biodegradation proceeds, this also offers an option for extending the bioremediation season into the winter months in cold climates. A potential drawback of composting is that it usually increases the volume of contaminated material, but if fully successful the finished compost can be returned to the site as a positive contribution to soil quality.

Slurry bioreactors offer the most aggressive approach to maximizing contact between the contaminated soil and the degrading organisms. Both lagoons and reactor vessels have been used, but the former are often not optimally designed for all the soil to be partially suspended by the mixing impellers. Contained reactor designs include mixing tank, airlift, and fluidized-bed aeration. A major advantage of contained slurry bioreactors is the potential ability to optimize nutrients, aeration and degradative inocula as fresh soil is added, and the control of waste materials, including gases continually. Slurry bioreactors are usually the most expensive bioremediation option because of the large power requirements, but under some conditions this cost is offset by the rapid biodegradation that can occur.

In all these cases it is important to bear in mind that although the majority of hydrocarbons are readily biodegraded, some, such as the steranes and hopanes, are very resistant to microbial attack. Estimates of oil biodegradation range from 60–95% for different crude oils, so fresh spills of crude oils are readily treated by bioremediation (21). Refined products, such as gasoline, diesel, jet fuels, and heating oils are usually more biodegradable than typical whole oils, but the various heavy fractions of crude oils, such as the asphalts, are far less biodegradable, and are not such attractive targets for bioremediation. Some crude oils have already been extensively biodegraded in their reservoirs, and these are also poor targets for bioremediation. An example is Orimulsion, a heavy oil in water emulsion (70% bitumen) stabilized by low levels of surfactants, used as a fuel for electricity generation. Similarly, old spills may have already undergone significant biodegradation and the residue may be relatively biologically inert. It is thus important to run laboratory studies to ensure that the contaminant is sufficiently biodegradable that clean-up targets can be met.

Halogenated Organic Solvents.

Constituents. Halogenated organic solvents are widely used in metal processing, electronics, dry cleaning and paint, paper and textile manufacturing, and some representative examples are shown in Figure 6. These solvents have been used for more than fifty years, and unfortunately they are fairly widespread contaminants. Unlike the hydrocarbons, which usually float on water, the halogenated solvents typically have specific gravities greater than 1, and they generally sink to the bottom of any groundwater, and float on the bedrock. For this reason they are sometimes known as DNAPLs for dense nonaqueous phase liquids.

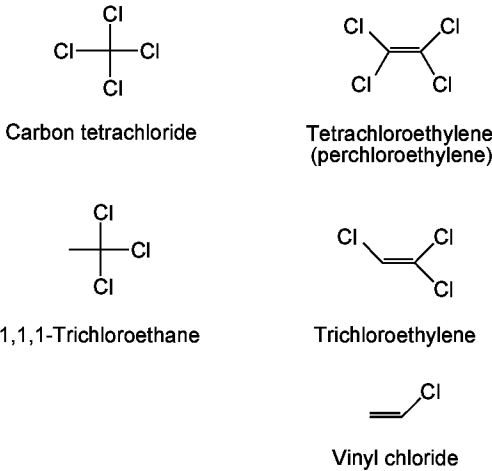


Fig. 6. Some representative halogenated solvents.

Biodegradation. Halogenated solvents are degraded under aerobic and anaerobic conditions. The anaerobic process is typically a reductive dechlorination that progressively removes one halide at a time (Fig. 7). For example, under methanogenic conditions, carbon tetrachloride is sequentially dechlorinated to chloroform, dichloromethane, methyl chloride and methane, while trichloroethylene is sequentially reduced to ethylene. Some of these compounds are also dehalogenated under sulfate-reducing conditions, and under denitrifying conditions there are reports that the final product can be CO₂ (22). Chloromethane and dichloromethane have been shown to be the sole carbon source for several anaerobic organisms (23), and it seems there is much to be learned about the microbial diversity of anaerobic microorganisms capable of dechlorinating solvents.

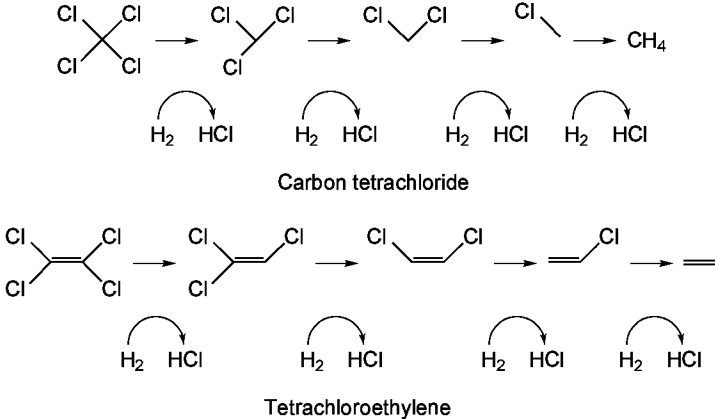


Fig. 7. Reductive dechlorination of carbon tetrachloride and tetrachloroethylene.

The simplest chlorinated alkanes, alkenes, and alcohols (eg, chloromethane, dichloromethane, chloroethane, 1,2-dichloroethane, vinyl chloride, and 2-chloro-ethanol) serve as substrates for aerobic growth for some bacteria, but the majority of halogenated solvents cannot support growth (24). Nevertheless these compounds are mineralized under aerobic conditions, albeit with no apparent benefit to the degrading organism. Indeed, the oxidation appears to be fortuitous and it occurs during the metabolism of a growth substrate. The phenomenon is therefore known as co-metabolism or co-oxidation. Numerous bacteria are able to catalyze the oxidation of trichloroethylene; some use monooxygenases (for example methane and ammonia oxidizing species); others contain dioxygenases (eg, some toluene oxidizing species). The difference between these two classes of enzymes is the fate of the two atoms of molecular oxygen. Monooxygenases insert one oxygen atom into their substrate, and reduce the other to water. Dioxygenases insert both atoms into their substrate. The effect of these two types of enzyme is illustrated in Figure 8. In either case, biodegradation proceeds to complete mineralization (25).

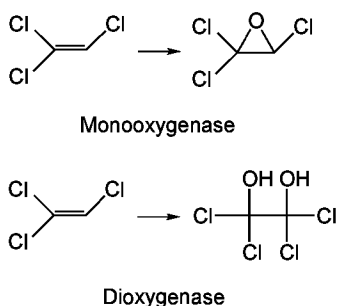


Fig. 8. Aerobic activation of tetrachloroethylene.

The biodegradation of trichloroethylene is the most studied since this is probably the most widespread halogenated solvent contaminant. Several substrates drive trichloroethylene co-oxidation, including methane, propane, propylene, toluene, isopropylbenzene, and ammonia (25). The enzymes that metabolize these substrates have subtly different selectivities with regard to the halogenated solvents, and to date none are capable of co-oxidizing carbon tetrachloride or tetrachloroethylene. Complete mineralization of these compounds can, however, be achieved by sequential anaerobic and aerobic process.

Bioremediation.

Air. Biofilters are an effective way of dealing with air from industrial processes that use halogenated solvents such chloromethane, dichloromethane, chloroethane, 1,2-dichloroethane and vinyl chloride, that support aerobic growth (26). Both compost-based dry systems and trickling filter wet systems are in use. Similar filters could be incorporated into pump-and-treat operations.

Groundwater and Soil. Halogenated solvents have contaminated soils and groundwater throughout the industrialized world, and remediation has a high priority. Since the solvents are so dense, they are typically found on the bedrock underlying aquifers. Pumping out the liquid phase is an obvious first step if the contaminant is likely to be mobile, but *in situ* bioremediation is a promising option. Thus, the U.S. Department of Energy is investigating the use of anaerobic *in situ* degradation of carbon tetrachloride in an aquifer some 76 m below the Hanford, Washington site, with nitrate as electron acceptor, and acetate as electron donor (22).

Trichloroethylene is the most frequent target of remediation, and as discussed above, this is only metabolized cometabolically. Remediation operations thus incorporate the addition of the cometabolized substrate. Methane was used successfully at the U.S. Department of Energy site at Savannah River, near Aiken, South Carolina (27) which had both an air-stripping and a biological component. Horizontal wells were used to pump methane and air below the contaminant, while an upper horizontal well in the vadose zone was used to withdraw these gases through the contaminated zone. Optimum biodegradation performance seemed to come from alternate injection of air and methane in air, and the inclusion of nitrous oxide and triethylphosphate, both gases, to give a C:N:P ratio of 100:10:1.

Plants may have a role to play in enhancing microbial biodegradation of halogenated solvents, for it has recently been shown that mineralization of radiolabelled trichloroethylene is substantially greater in vegetated rather than unvegetated soils (28), indicating that the rhizosphere provides a favorable environment for microbial degradation of organic compounds.

Methane has also been used in aerobic bioreactors that are part of a pump-and-treat operation, and toluene and phenol have also been used as co-substrates at the pilot scale (29). Anaerobic reactors have also been developed for treating trichloroethylene. For example, Wu and co-workers (30) have developed a successful upflow anaerobic methanogenic bioreactor that converts trichloroethylene and several other halogenated compounds to ethylene.

Groundwater contaminated with other halogenated solvents can also be treated in aboveground reactors. Aerobic reactors are useful for those compounds that can support growth. For example, a membrane reactor has been designed for treating 1,2-dichloroethane (31), and bubble columns and packed bed reactors have been developed for the aerobic degradation of 2-chloroethanol (32). As mentioned above, sequential anaerobic and aerobic reactors are capable of mineralizing tetrachloroethylene (33).

Halogenated Organic Compounds.

Constituents. Complex halogenated organic compounds have been widely used in commerce in the last fifty years. A few representative examples are shown in Figure 9; pentachlorophenol has been widely used as a wood preservative, and also for termite control. (2,4-Dichlorophenoxy)acetate (2,4-D) is widely used as a broad-leaf herbicide, DDT was widely used as an insecticide, and hexachlorophene has been widely used as a germicide. Polychlorinated biphenyls (PCBs) were sold with varying levels of chlorination for a range of purposes. They ranged from light oily fluids (with two, three, or four chlorines) to viscous oils (five chlorines) to greases and waxes (six or more chlorines), and their names indicated the level of chlorination. Thus Aroclor 1242 (trademark of Monsanto, U.S.), Clophen A30 (trademark of Farbenfabriken Bayer AG, Germany) and Kanechlor 300 (trademark of Kane-gafuchi Chemical Industries, Japan) all contained 42% chlorine by weight, and an average of three chlorines per biphenyl. An important property shared by all these compounds is their relative resistance to biodegradation, so at first glance they may not seem a good target for bioremediation. Indeed complex halogenated organic compounds were widely thought to be almost exclusively anthropogenic in origin, so that there would have been little time for biodegradation pathways to evolve. This view is being corrected, for in fact a variety of organisms, particularly marine algae and some fungi, produce significant quantities of these compounds (34). There is, thus, good reason to expect that halogenated-organic degrading organisms will be found in the biosphere, and this has been borne out in practice.

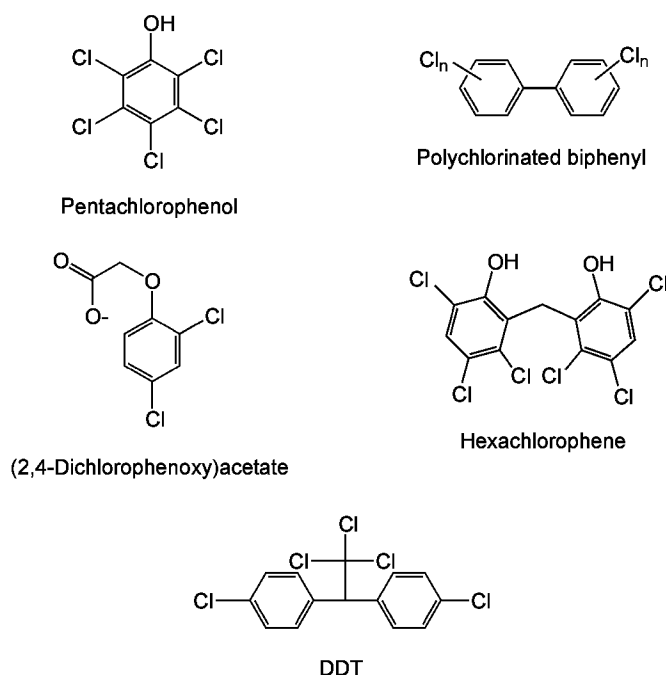


Fig. 9. Representative halogenated organic contaminants.

Biodegradation. An important characteristic of degradation is the cleavage of carbon–chlorine bonds, and the enzymes that catalyze these reactions, the dehalogenases, are being characterized (35). The reductive dechlorination seen with carbon tetrachloride and tetrachloroethylene (see Fig. 7) seems to be a general phenomenon, and even compounds as persistent as DDT and the polychlorinated biphenyls are reductively dechlorinated under some conditions, particularly under methanogenic conditions. Some compounds, such as pentachlorophenol, can be completely mineralized under anaerobic conditions, but the more recalcitrant ones require aerobic degradation after reductive dehalogenation.

Pentachlorophenol can be mineralized aerobically and anaerobically, and both processes have been exploited for bioremediation. Under methanogenic conditions a reductive dehalogenation, analogous to that seen with halogenated solvents, eventually generates phenol, although some of the intermediate congeners are quite recalcitrant (36). The phenol is further mineralized to methane and carbon dioxide, and sulfate-reducing bacteria may be involved (37).

Several bacterial isolates are able to grow aerobically using pentachlorophenol as sole source of carbon and energy, but many grow rather better when supplemented with a more nutritious substrate. Unfortunately, in many cases it seems that the more vigorous growth with these substrates does not enhance the biodegradation of pentachlorophenol (38). The initial step in pentachlorophenol biodegradation in one *Flavobacterium* species is an NADPH-dependent oxygenolytic dechlorination, where the *para*-chloro group is replaced by a hydroxyl to generate tetrachlorohydroquinone. Other species seem to produce the same metabolite by a hydrolytic process (38). Fungal degradation of pentachlorophenol, apparently using the lignolytic apparatus, has also been reported (39).

(2,4-Dichlorophenoxy)acetate (2,4-D) has been one of the world's most popular herbicides. Although it is somewhat resistant to biodegradation, it is biodegraded by several bacterial isolates. It is a general truism that the more halogens on a molecule, the slower its biodegradation, and this is borne out with the related herbicide 2,4,5-T ((2,4,5-trichlorophenoxy)acetate). Nevertheless, bacterial degradation has been seen under both aerobic and anaerobic conditions, the latter involving reductive dechlorination via 2,4-D. Aerobic degradation removes acetate from (2,4-dichlorophenoxy)acetate to yield 2,4-dichlorophenol, which is subsequently hydroxylated to 3,5-dichlorocatechol, followed by ring cleavage and complete mineralization (40). Genetic engineering has been used to construct strains that are particularly adept at consuming these compounds, but whether these will overcome the regulatory hurdles to allow their use outside the laboratory remains to be seen.

DDT (1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane) is a remarkably resistant molecule, which explains both its efficacy as an insecticide, and its accumulation at the top of the food-chain. Nevertheless, there are indications that it can be biodegraded, both anaerobically with initial dechlorination and aerobically with initial ring hydroxylation (40). DDT and its partially degraded congeners are very hydrophobic, and biodegradation seems to be stimulated by adding surfactants. White-rot fungi also degrade DDT under lignolytic conditions, although there is little mineralization to CO₂ (5).

The polychlorinated biphenyls are quite recalcitrant. Some lightly chlorinated biphenyls are readily mineralized under aerobic conditions (41), and indeed the structure of an enzyme that catalyzes the key ring-cleavage oxygenation has recently been determined by x-ray crystallography (42). More chlorinated congeners are resistant to aerobic degradation, but they are reductively dechlorinated under anaerobic conditions. Complete degradation of the commercial mixtures, thus, generally requires an anaerobic process followed by an aerobic one. A major issue for the oxidative process is that biphenyl seems to be required for significant expression of the biodegradative system (41); the chlorinated compounds do not induce the enzymes that would degrade them. The biochemistry of the biodegradative process is only beginning to be unraveled, but already there are suggestions that there is considerable diversity in the enzyme systems able to degrade these compounds. There is a lot of effort aimed at engineering organisms to degrade polychlorinated biphenyls, and in finding ways to make these very hydrophobic compounds more bioavailable by the use of chemical oxidants, such as Fenton's reagent (43), or surfactants.

Bioremediation.

Soil. Pentachlorophenol has been the target of bioremediation at a number of wood-treatment facilities, and good success has been achieved in several applications. It is rarely the sole contaminant, and is often present with polynuclear hydrocarbons from coal creosote. *In situ* degradation has been stimulated by bioventing, where air is injected through some wells, and extracted through others to both strip volatiles and provide oxygen to indigenous organisms (44). Just as with the halogenated solvents, it seems that plants stimulate microbial degradation of pentachlorophenol in the rhizosphere (45).

The kinetics of such *in situ* degradation are rather slow, however, and more active bioremediation is usually attempted. For example, contaminated soil at the Champion Superfund site in Libby, Montana, was placed into 1-acre land treatment units in 6-in. layers, and irrigated, tilled, and fertilized. Under these conditions, the half-lives of pentachlorophenol, pyrene, and several other polynuclear aromatic hydrocarbons, initially present at around 100–200 ppm, were on the order of 40 days (46). This success relied on the indigenous microbial populations in the soil, but many groups are focusing on the addition of organisms (eg, 47), perhaps immobilized on some sort of carrier. Composting, and bioremediation focusing on the use of white-rot fungi, has also met with success at the pilot scale. Others have used fed-batch or fluidized-bed bioreactors to stimulate the biodegradation of pentachlorophenol. This allows significant optimization of the process and increases in rates of degradation by tenfold (48).

A major concern when remediating wood-treatment sites is that pentachlorophenol was often used in combination with metal salts, and these compounds, such as chromated copper–arsenate, are potent inhibitors of at least some pentachlorophenol degrading organisms (49). Sites with significant levels of such inorganics may not be suitable candidates for bioremediation.

The phenoxy-herbicide, 2,4-D, has been successfully bioremediated in a soil contaminated with such a high level of the compound (710 ppm) that it

was toxic to microorganisms (50). There were essentially no indigenous bacteria in the soil. Success relied on washing a significant fraction of the contaminant off the soil and adding bacteria enriched from a less contaminated site. Success was achieved in remediating both soil washwater and soil in a bioslurry reactor (50). 2,4-D is also effectively degraded in composting, with about half being completely mineralized, and the other half becoming incorporated in a nonextractable form in the residual soil organic matter (51).

The bioremediation of polychlorinated biphenyls in soils is receiving significant attention because these compounds are quite widely distributed in the environment, either from leaking electrical transformers or sometimes because they were applied as part of road maintenance. In the latter case, the contamination usually includes petroleum hydrocarbons, and unfortunately it seems that the two contaminants inhibit the degradation of each other. Nevertheless, cultures are being found that can degrade both polychlorinated biphenyls and petroleum hydrocarbons. There is also interest in the role of rhizosphere organisms in polychlorinated biphenyl degradation, particularly since some plants exude phenolic compounds into the rhizosphere that can stimulate the aerobic degradation of the less chlorinated biphenyls.

Groundwater. A successful groundwater bioremediation of pentachlorophenol is being carried out at the Libby Superfund site described above. A shallow aquifer is present at 5.5 to 21 m below the surface, and a contaminant plume is nearly 1.6 km in length. Nutrients and hydrogen peroxide were added at the source area and approximately half way along the plume, and pentachlorophenol concentrations decreased from 420 ppm to 3 ppm where oxygen concentrations were successfully raised. A membrane oxygen dissolution system has now been installed to replace the hydrogen peroxide additions, and costs have been substantially lowered without an apparent decrease in remediation performance (52).

Pentachlorophenol is readily degraded in biofilm reactors (53), so bioremediation is a promising option for the treatment of contaminated groundwater brought to the surface as part of a pump-and-treat operation.

River and Pond Sediments. Much of the work on polychlorinated biphenyls has focused on the remediation of aquatic sediments, particularly from rivers, estuaries, and ponds. As noted above, a few of the most lightly chlorinated compounds are mineralized under aerobic conditions, but the more chlorinated species seem completely resistant to aerobic degradation, even by white rot fungi (54). On the other hand, there is extensive dechlorination of highly chlorinated forms under anaerobic conditions, particularly methanogenic conditions. Bioremediation thus requires anaerobic and aerobic regimes. Intrinsic biodegradation of polychlorinated biphenyls can be recognized by the changing "fingerprint" of the individual isomers as biodegradation proceeds (41,55). The anaerobic dechlorination of the most recalcitrant congeners can apparently be primed by adding a readily dehalogenated congener, such as 2,5,3',4'-tetrachlorobiphenyl (56), but whether this is a realistic approach for *in situ* bioremediation remains to be seen. Harkness and co-workers (57) have successfully stimulated aerobic biodegradation in large caissons in the Hudson River by adding inorganic nutrients, biphenyl, and hydrogen peroxide, but found that repeated addition of a polychlorinated-biphenyl degrading bacterium (*Alcaligenes eutrophus* H850) had no beneficial effect. Essentially no biodegradation occurred in the stirred control caissons, but losses on the order of 40% were seen in the caissons that received nutrients and peroxide, regardless of whether the stirring was aggressive or rather gentle. Whether this approach can be scaled-up for large-scale use, with a net environmental benefit, remains to be seen.

Nonchlorinated Pesticides and Herbicides.

Constituents. A vast number of compounds are used as herbicides, pesticides, fungicides, etc, and a few are shown in Figure 10. In order to be effective they must have their effect before they are degraded in the environment, but on the other hand they must not be so resistant to degradation that they accumulate where they are used, or accumulate in food chains. It is unusual for these compounds to become contaminants where they are applied correctly, but manufacturing facilities, storage depots and rural airfields where crop-dusters are based have had spills that can lead to long lasting contamination. Bioremediation is a promising technology to remediate such sites. There are also some locations where groundwater has become contaminated by these chemicals, and again bioremediation may be a cost-effective remediation strategy.

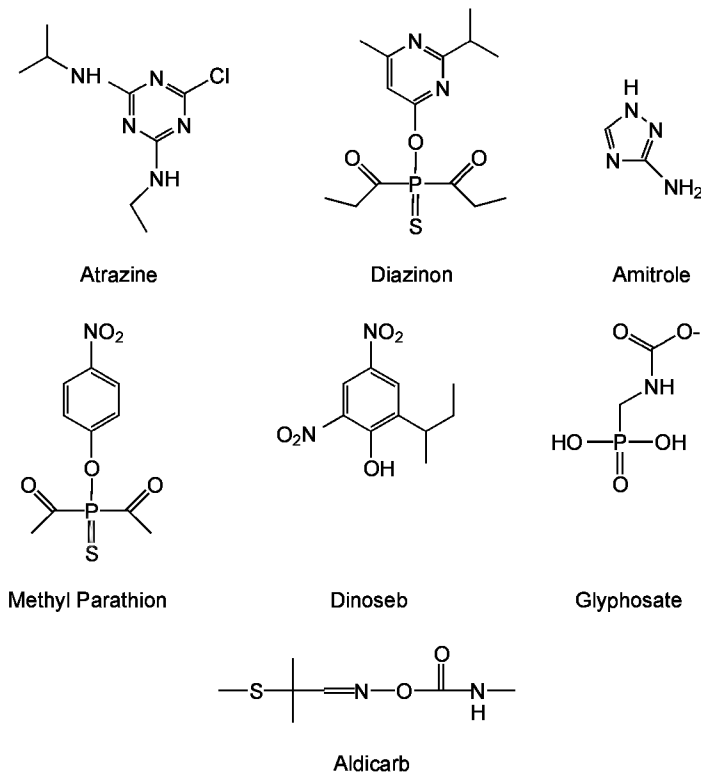


Fig. 10. Examples of herbicides and pesticides amenable to bioremediation.

Biodegradation. The vast majority of pesticides, herbicides, fungicides, and insecticides in use today are biodegradable, although the intrinsic biodegradability of individual compounds is one of the variables used in deciding which compound to use for which task. Some herbicides are acutely toxic to plants that absorb them, but are so readily degraded in soil that seeds can be planted at the same time as the herbicide application. Other herbicides are known to be effective at preventing plant growth for many months, and are used in situations where this is the desired goal. Very few degradation pathways of these compounds have been worked out in detail, but some generalizations can be made.

Compounds with organophosphate moieties, such as Diazinon, Methyl Parathion, Coumaphos and Glyphosate are usually hydrolyzed at the phosphorus atom (40,58). Indeed several *Flavobacterium* isolates are able to grow using parathion and diazinon as sole sources of carbon.

Triazines pose rather more of a problem, probably because the carbons are in an effectively oxidized state so that no metabolic energy is obtained by their metabolism. Very few pure cultures of microorganisms are able to degrade triazines such as Atrazine, although some *Pseudomonads* are able to use the compound as sole source of nitrogen in the presence of citrate or other simple carbon substrates. The initial reactions seem to be the removal of the ethyl or isopropyl substituents on the ring (41), followed by complete mineralization of the triazine ring.

Nitroaromatic compounds, such as Dinoseb, are degraded under aerobic and anaerobic conditions (59). The nitro group may be cleaved from the molecule as nitrite, or reduced to an amino group under either aerobic or anaerobic conditions. Alternatively, the ring may be the subject of reductive attack. Thus, while these molecules are sometimes quite long-lived in the environment, they can be completely mineralized under appropriate conditions (59). Recent work has isolated a *Clostridium bifermentans* able to anaerobically degrade dinoseb cometabolically in the presence of a fermentable substrate. The dinoseb was degraded to below detectable levels, although only a small fraction was actually mineralized to CO₂ (60).

Carbamates such as Aldicarb undergo degradation under both aerobic and anaerobic conditions. Indeed the oxidation of the sulfur moiety to the sulfoxide and sulfone is part of the activation of the compound to its most potent form. Subsequent aerobic metabolism can completely mineralize the compound, although this process is usually relatively slow so that it is an effective insecticide, acaricide and nematocide. Anaerobically these compounds are hydrolyzed, and then mineralized by methanogens (61).

Bioremediation.

Groundwater. Atrazine dominated the world herbicide market in the 1980s, and contamination of groundwater has been reported in several locations in the U.S., Europe, and South Africa. There are several reports that once in groundwater it is very recalcitrant, suggesting that atrazine-degrading organisms are not widespread. Nevertheless, successful biodegradation has been achieved with indigenous organisms in laboratory mesocosms after a lag phase, and once activity was found, it remained (62). Interestingly the degradation was somewhat slowed by the addition of low concentrations of readily assimilate carbon, such as lactate, and it is not clear how biodegradation might be stimulated in the field. Nevertheless, it is clear that intrinsic remediation is likely to lead to the disappearance of atrazine from groundwaters once atrazine-utilizers have become abundant, and perhaps inoculation with atrazine-metabolizing species will be effective.

If more active treatment is required, such as pump-and-treat, it is possible that biological reactors will be a cost-effective replacement for activated carbon filters (63).

Marsh and Pond Sediments. Herbicides and pesticides are detectable in marsh and pond sediments, but intrinsic biodegradation is usually found to be occurring. Little work has yet been presented where the biodegradation of these compounds has been successfully stimulated by a bioremediation approach.

Soil. Herbicides and pesticides are of course metabolized in the soil to which they are applied, and there are many reports of isolating degrading organisms from such sites. Degradation activities are typically much higher at sites that have seen product application, indicating that natural enrichment processes occur. Much current effort is aimed at assessing the diversity of degradative pathways, and in many cases it seems that several different natural metabolic pathways can degrade individual pollutants. Little work has yet been presented where the biodegradation of these compounds has been successfully stimulated by a bioremediation approach, but inoculation with active organisms may be a promising approach (64,65).

It is a general observation that herbicide degradation occurs more readily in cultivated than fallow soil, suggesting that rhizosphere organisms are effective herbicide degraders. Whether this can be effectively exploited in a phytoremediation strategy remains to be seen.

Military Chemicals.

Constituents. The military use a range of chemicals as explosives and propellants, which are sometimes termed "energetic molecules". Generally speaking, modern explosives are cyclic, often heterocyclic, composed of carbon, nitrogen and oxygen. Perhaps the most well known is 2,4,6-trinitrotoluene (Fig. 11), but RDX (Royal Demolition eXplosive; hexahydro-1,3,5-trinitro-1,2,3-triazine) and HMX (High Melting eXplosive; octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine), are even more powerful. *N,N*-Dimethylhydrazine is used as a solid rocket fuel. These compounds are sometimes present at quite high levels in soils and groundwater on military bases and production sites. One quite infamous problem at the latter is "pink water", a relatively undefined mixture of photodegradation products of TNT. Bioremediation is a promising new technology for treating sites contaminated with such compounds.

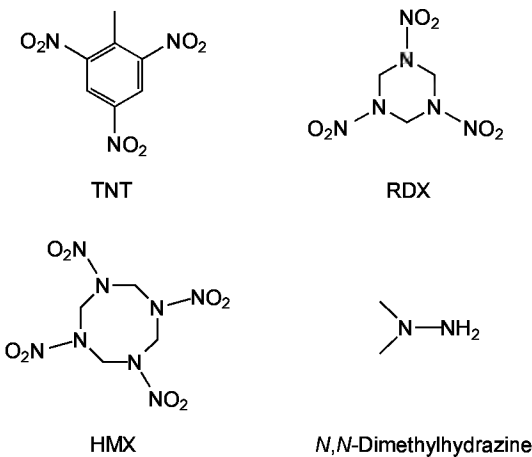


Fig. 11. Military explosives and a rocket propellant.

Bioremediation may also be an appropriate tool for dealing with chemical agents such as the mustards and organophosphate neurotoxins (66), but little work on actual bioremediation has been published.

Biodegradation. Natural nitrosubstituted organic compounds are quite unusual, and it was once thought that their degradation was principally by abiotic processes. As shown above in the case of Dinoseb, however, nitrosubstituted compounds are subject to a variety of degradative processes. The biodegradation of TNT is well established (67). Under anaerobic conditions it is readily reduced to the corresponding aromatic amines and subsequently deaminated to toluene. As shown in the section on hydrocarbons, the latter can be mineralized under anaerobic conditions, leading to the potentially complete mineralization of TNT in the absence of oxygen.

Under aerobic conditions TNT can be mineralized by a range of bacteria and fungi, often co-metabolically with the degradation of a more degradable substrate. There is even evidence that some plants are able to deaminate TNT reductively. However, there is also ample evidence that under some conditions, the TNT is converted to insoluble large molecular weight compounds. This probably occurs by addition reactions to soil components, such as humic and fulvic acids, or cellular material in the case of plants. Lignolytic fungi are also yielding promising results for the degradation of both TNT and "pink water", particularly if the latter is pretreated with uv irradiation.

RDX and HMX are rather more recalcitrant, especially under aerobic conditions, but there are promising indications that biodegradation can occur under some conditions, especially composting (67). Several strains of bacteria able to use RDX (and Triazine) as a sole source of nitrogen for growth have recently been isolated, and this is an area where rapid progress is being made.

Little work has been reported on the biodegradation of dimethylhydrazine, but it may become an important target for remediation at some sites (68).

Bioremediation.

Groundwater. Nitrotoluenes have been detected in groundwater in some areas, and intrinsic remediation may be occurring at some sites by anaerobic degradation. Research into whether this can be stimulated with a net environmental benefit is in its very earliest stages, and no clear evidence for success has been presented.

A commercial technology (69), the SABRE process, treats contaminated water and soil in a two-stage process by adding a readily degradable carbon and an inoculum of anaerobic bacteria able to degrade the contaminant. An initial aerobic fermentation removes oxygen so that the subsequent reduction of the contaminant is not accompanied by oxidative polymerization.

Soil. Composting of soils contaminated by high explosives is being carried out at the Umatilla Army Depot near Hermiston, Oregon (70). Soil from munitions washout lagoons is being treated indoors in compost rows of 2,000 m³, and the estimated cost is less than one-third the estimated cost of incineration. If this is successful, there are 30 similar sites on the National Priority List that could be treated in a similar way.

Other Organic Compounds. The majority of organic compounds in commerce are biodegradable, so bioremediation is a potential option for cleaning up after industrial and transportation accidents. For example, Ref. 71 reports the successful bioremediation of 23,000 liters of vinyl acetate spilled from a railroad tank car in Albany, New York, at a cost approximately half that of excavation and disposal of the 1,100 m³ of contaminated soil. They also report that *in situ* biological treatment has been used to remediate spills of other organics, including acrylonitrile, styrene, 2-butoxyethanol, and ethacrylate, at other sites where there had been railroad accidents. Bioremediation is thus already an important tool in remediating accidental spills of organic compounds.

Bioremediation is also an option when spills of such compounds contaminate groundwater. For example, bioremediation seems a feasible treatment for aquifers contaminated with alkyipyridines (72) and phenol (73).

Inorganic Contaminants

Nitrogen Compounds.

Constituents. Nitrogen-containing compounds are of concern for several reasons. Nitrate levels are regulated in groundwater because of concerns for human and animal health. Ammonia is regulated in streams and effluents as a potential fish toxicant, and any nitrogenous contaminant is a potential problem in water because of its stimulatory effect on the growth of algae. Other nitrogenous contaminants include cyanides in mine waters. Fortunately, all are amenable to biological treatment.

Biodegradation. The biological mineralization of fixed nitrogen is well studied; ammonia is oxidized to nitrite, and nitrite to nitrate, by autotrophic bacteria, and nitrate is reduced to nitrogen by anaerobic bacteria. Urea in sewage and industrial wastes is readily hydrolyzed to ammonia and CO₂ by many bacteria, and cyanides and cyanates are used as sole sources of carbon and nitrogen by some organisms. Wastewater treatment facilities utilize these organisms in assuring that municipal and industrial effluents meet strict water quality standards, but this biological process is outside the scope of this article (see WATER, INDUSTRIAL WATER TREATMENT; WATER, MUNICIPAL WATER TREATMENT). On the other hand, ammonia and nitrate are essential nutrients for plant and bacterial growth, so one option is to use these organisms to take up and use the contaminants.

Bioremediation.

Surface Water. One example of exploiting biology to handle excess nitrogen in a surface water is the case of the Venice Lagoon in Italy. About a million tons of sea lettuce (*Ulva*) grows in the lagoon annually because of the high levels of nitrogenous nutrients in this relatively landlocked bay. This material is harvested, composted and sold as a low-cost remediation of this problem (74). In areas where water quality is sufficiently high, growing even more valuable crops, such as nori (*Porphyra*) may be an attractive bioremediation option.

A more constrained opportunity for nitrate bioremediation arose at the US-DoE Weldon Spring Site near St. Louis, Missouri. This site had been a uranium and thorium processing facility, and treatment of the metal had involved nitric acid. The wastestream, known as raffinate, was discharged to surface inpoundments and neutralized with lime to precipitate the metals. Two pits had nitrate levels that required treatment before discharge, but heavy rains in 1993 threatened to cause the pits to overflow. Bioremediation by the addition of calcium acetate as a carbon source successfully treated more than 19 million liters of water at a reasonable cost (75).

Groundwater. One approach to minimizing the environmental impact of excess nitrogen in groundwater migrating into rivers and aquifers is to intercept the water with rapidly growing trees, such as poplars, that will use the contaminant as a fertilizer.

An alternative approach is to add a readily degradable substrate to the contaminated aquifer, in the absence of oxygen, to stimulate bacterial denitrification. Soluble substrates such as ethanol have been suggested, but a recent alternative suggestion is to provide vegetable oil, either down the well itself, or on a gravel trickling filter. Since the oil is insoluble in water, it becomes immobilized on the aquifer material, allowing denitrification as the water flows through (76).

Metals and Metalloids. A wide range of metallic and nonmetallic elements are present as contaminants at industrial and agricultural sites throughout the world, both in ground- and surface water, and in soils. They pose a quite different problem from that of organic contaminants, since they cannot be degraded so that they disappear. Some metal and metalloid elements have radically different bioavailabilities and toxicities depending on their redox state, so one option is stabilization by converting them to their least toxic form. This can be a very effective way of minimizing the environmental impact of a contaminant, but if the contaminant is not removed from the environment, there is always the possibility that natural processes, biological or abiological, may reverse the process. Removing the contaminants from water phases is relatively straightforward, and the wastewater treatment industry practices this on an enormous scale. Pump-and-treat systems that mimic waste water treatments are already being used for several contaminants, and less complex systems involving biological mats are a promising solution for less demanding situations.

Microbial leaching of metals from ores is a promising adjunct to more aggressive metal recovery technologies (77), but is generally achieved by oxidative processes that generate very acidic waters. It seems unlikely that similar approaches will be of much value in removing contaminant metals and metalloids from soils.

In the past, removing metal and metalloid contaminants from soil has been impossible, and site clean-up has meant excavation and disposal in a secure landfill. An exciting new approach to this problem is phytoextraction, where plants are used to extract contaminants from the soil and harvested.

Immobilization and Toxicity-Minimization.

Adsorption. Biomass, often agricultural by-products, has been widely used as an adsorbent for metals and other contaminants in water (78), but this is outside the scope of this article.

Microbial Processes. Most elements display a range of solubilities and biological effects depending on their chemical form. For example, chromium, although it may be an essential micronutrient for many organisms, is known to be toxic at higher levels. Indeed chromium has a range of available redox states that differ significantly in their environmental impact. Cr(VI) is generally more soluble and more toxic than Cr(III), but a wide range of organisms, both bacteria and fungi, are able to reduce the former to the latter under both aerobic and anaerobic conditions. Thereby the environmental impact of the contaminant is reduced. Since Cr(III) precipitates as insoluble hydroxides under neutral and alkaline conditions, microbial reduction is a potential treatment for soils and waters (79).

Similarly selenium is a micronutrient that is toxic at higher levels. It also has quite different bioavailability in its different redox states, with the elemental form being the least biologically available. Many microorganisms, both bacteria and fungi, are able to reduce more oxidized species, especially selenite (Se(IV)) to the red elemental form, providing an appealing remediation option for this element. An alternative approach to remediating selenium contamination is to encourage methylation to volatile dimethylselenide (80). Although the microbiology of this process is not yet very well understood, it seems to be the result of degradation of the selenium-containing amino acid, selenomethionine.

Arsenic is another element with different bioavailability in its different redox states. Arsenic is not known to be an essential nutrient for eukaryotes, but arsenate (As(V)) and arsenite (As(III)) are toxic, with the latter being rather more so, at least to mammals. Nevertheless, some microorganisms grow at the expense of reducing arsenate to arsenite (81), while others are able to reduce these species to more reduced forms. In this case it is known that the element can be immobilized as an insoluble polymetallic sulfide by sulfate reducing bacteria, presumably adventitiously due to the production of hydrogen sulfide (82). Indeed many contaminant metal and metalloid ions can be immobilized as metal sulfides by sulfate reducing bacteria.

A rather more specific mechanism of microbial immobilization of metal ions is represented by the accumulation of uranium as an extracellular precipitate of hydrogen uranyl phosphate by a *Citrobacter* species (83). Staggering amounts of uranium can be precipitated; more than 900% of the bacterial dry weight! Recent work has shown that even elements that do not readily form insoluble phosphates, such as nickel and neptunium, may be incorporated into the uranyl phosphate crystallites (84). The precipitation is driven by the production of phosphate ions at the cell surface by an external phosphatase.

Although the process requires the addition of a phosphate donor, such as glycerol-2-phosphate, it may be a valuable tool for cleaning water contaminated with radionuclides. An alternative mode of uranium precipitation is driven by sulfate-reducing bacteria such as *Desulfovibrio desulfuricans*, which reduce U(VI) to insoluble U(IV). When combined with bicarbonate extraction of contaminated soil, this may provide an effective treatment for removing uranium from contaminated soil (85).

Microbial processes can also detoxify mercury ions and organic compounds by reducing the mercury to the elemental form, which is volatile (86). This certainly reduces the environmental impact of compounds such as methylmercury, however, such a bioprocess would have to include a mercury capture system before it could be exploited on a large scale with public support.

Rhizofiltration. Rhizofiltration is the use of plant root systems to remove contaminating metal ions from water (87). The plants must have substantial root systems, and success has been reported with both hydroponically grown terrestrial plants, such as sunflowers (*Helianthus annuus*), and floating aquatic plants such as water hyacinth (*Eichhornia crassipes*), and water milfoil (*Myriophyllum spicatum*). A filamentous cyanobacterium, *Phormidium* sp., has also proven quite effective at removing metals from water. Understanding rhizofiltration at the molecular level is at an early stage, but it is already clear that several processes are involved. Adsorption of the contaminant to the root surface must occur first, and presumably happens in all plants, probably to polygalacturonic acids for cations, and positively charged polypeptides for anions. Some plants actively translocate the metal ions into their cells and some transport the metal to the leafy shoots. This can be exploited in phytoextraction of metal ions from soils (see below). This phenomenon may actually be undesirable in rhizofiltration, because if the metals are retained in the roots there will be a smaller volume of contaminated material to treat at the end of the water treatment. Perhaps of more importance for rhizofiltration is the process of root assisted precipitation, where metal ions are precipitated on the roots as insoluble inorganic complexes. For example, lead is precipitated as lead phosphates on roots of Indian Mustard (*Brassica juncea*) to a remarkable level; up to 45% of dry weight (87). This phenomenon seems to require active exudation of precipitants from the roots, for it is dramatically reduced when dead roots are used.

Phytoextraction. The fact that certain plant species are able to survive on soils with such high levels of metal ions that the growth of most plants is inhibited has been known for centuries. However, it was not until this century that it became clear that some of these plants actually accumulated the potentially toxic metals and somehow resisted their toxic effects. These plants, known as hyperaccumulators, are exemplified by Alpine Pennycress, *Thlaspi caerulescens*, which is able to accumulate a wide range of metals, including cadmium, chromium, copper, lead, nickel, and zinc, more than several thousand fold in its roots, and somewhat less in its leaves (88). Unfortunately for a use in bioremediation, most hyperaccumulators are small, slow growing plants. Researchers have, therefore, turned to more rapidly growing plants which, while they might not be quite as effective as the hyperaccumulators on a weight basis, might be able to extract more metal in a growing season (88). Indian mustard, *Brassica juncea*, was found to be almost as effective as *T. caerulescens* and to grow much faster. This plant is now being used in field trials of phytoextraction.

Phytoextraction of metals from soils requires plants that have substantial root systems to maximize contact with the contaminated soil, and effective transport mechanisms to get the metal from the root to harvestable biomass. Merely accumulating the metal in or on the roots would be less desirable, because harvesting would likely increase the environmental impact of the contaminant by creating dust. It has been known for some time that metals such as cadmium are often stored in plant tissues as phytochelatin-complexes. Phytochelatins are small peptides rich in cysteines that chelate metals via the cysteinyl sulfur (89). Other elements are stored as metalphytate precipitates (88), and in both cases seem to be concentrated in the vacuolar sap. But how the metals are transported in the xylem sap from the root to the shoot is only beginning to be explored. Cadmium seems to be transported as a soluble salt of organic acids (90), at least in *B. juncea*, while nickel is thought to be transported as a soluble histidine complex in *Alyssum* species (91). Optimizing this process will be an important consideration for the commercial success of phytoremediation.

It will also be important to understand the rhizosphere ecology around the roots of metal accumulating plants fully. Maximizing the bioavailability of the contaminant metals in this zone may require the optimization of the microbial communities, or perhaps the addition of soil amendments. There are early indications that such intervention may be beneficial (88), but research in this area is at a very early stage.

Phytoremediation is also being developed for dealing with soils contaminated with high levels of selenium in California; again *B. juncea* seems to be particularly effective in accumulating the contaminant from soil, and all plants tested were more effective at removing selenate than selenite (92). This is an interesting contrast to bacterial systems, where selenite reduction is more commonly found than selenate reduction.

There have also been reports that some plants, including *B. juncea*, stimulate volatilization of dimethylselenide, although it seems that this is an indirect effect where the plant roots stimulate bacterial evolution of the gas (93). Presumably the selenium is eventually oxidized in the atmosphere and returned to the soil as rain. Since much of the world is marginally selenium deficient, such a process might have no deleterious environmental effects. A similar volatilization approach may eventually be used for mercury contamination. Ref. 94 describes the effective transfer of bacterial mercury reduction genes into a plant. In the future, this approach might offer an option for cleaning mercury-contaminated soils, albeit with some form of mercury capture technology.

Phytostabilization. In some cases it may be important to stabilize a contaminated site to minimize harmful environmental impacts. Barren soils are more prone to erosion, leaching, and dust formation than vegetated soils, so establishing plants as a ground cover may be a very beneficial treatment, even if the contaminant remains in place. In some cases, planting metal tolerant species is enough to establish a plant community, but in other situations, such as mine waste piles, it may be essential to restore fertility to the soil by adjusting the pH and adding organic composts. Where phytoextraction requires plants that actively transport metals to the above-ground biomass, this would be an undesirable phenomenon in phytostabilization programs because it would promote the movement of the contaminating metals into the biosphere.

It may also be possible to stimulate bioreductions in soils, such as Cr(VI) to Cr(III), more effectively by growing plants than by adding bacteria or nutrients to stimulate the microbial process.

Bioremediation.

Water. Groundwater can be treated in anaerobic bioreactors that encourage the growth of sulfate reducing bacteria, where the metals are reduced to insoluble sulfides, and concentrated in the sludge. For example, such a system is in use to decontaminate a zinc smelter site in the Netherlands (95).

Bacterial remediation of selenium oxyanions in San Joaquin, California, drainage water is under active investigation (96,97), but has not yet been commercialized. Agricultural drainage rich in selenium is also typically rich in nitrates, so bioremediation must also include conditions that stimulate denitrification (98). Phytoextraction of selenium is also being tested, but is not yet being used on a large scale.

Phytoremediation is also not yet being used commercially, but results at several field trial suggest that commercialization is not far away. Perhaps the biggest success to date is the successful rhizofiltration of radionuclides from a Department of Energy site at Ashtabula, Ohio, where uranium concentrations of 350 ppb were reduced to less than 5 ppb, well below groundwater standards, by Sunflower roots (99). Small field tests also suggest that rhizofiltration will be a remediation option for removing radionuclides, such as ¹³⁷Cs and ⁹⁰Sr, from contaminated lakes and ponds near the Chernobyl nuclear reactor in Ukraine.

Mine Drainage. Natural drainage water that come into contact with active and abandoned metal and coal mines can become seriously contaminated with a range of heavy-metal ions, and or often become quite acidic, with a pH near 2. While underground, the water is typically anoxic, and any iron is present as soluble ferrous species. When this mixes with aerobic surface water, the iron precipitates as bright orange ferric hydroxide, and this can have a serious environmental impact. In recent years it has become clear that the environmental impact of acid mine drainage can be minimized by the construction of artificial wetlands that combine geochemistry and biological treatments. These systems are being designed for a range of wastewaters, most of which fall outside the scope of this article.

The precipitation of ferric hydroxide is typically biologically mediated by iron-oxidizing bacteria at acid pH, but is usually rather slow. Abiological oxidation becomes more important at pH values above 5, and this is usually much faster than the biological process. Most constructed wetlands for treating acid mine waters thus start with a zone designed to raise the pH. A bed of crushed limestone often suffices to raise the pH significantly, and it is important that this be kept anoxic to prevent rust precipitation on the limestone, which would prevent further production of alkalinity. Once the pH is near neutral, the water is discharged into an aerobic wetland to encourage the precipitation of iron and aluminum oxides, and the co-precipitation of arsenic, if this is present (100,101).

If heavy metals are present in the mine water, the iron-free water can be made to flow into an anaerobic part of the constructed wetland, where organic material, such as compost, manure or sawdust, provides reductants to sulfate-reducing bacteria that become established. These bacteria reduce sulfate to sulfide, which precipitates the heavy metal ions as insoluble sulfides. It is important that this part of the constructed wetland be kept anaerobic, to prevent oxidation and remobilization of the precipitated metals. For this reason, this part of the wetland is typically kept flooded and free of aquatic plants that might introduce oxygen through their roots (100). Finally, an aerobic algal mat can act as a polishing step to complete the removal of contaminants, particularly manganese (101,102). Mine drainage that is not acidic may not need such complicated systems and individual parts of the treatment train described above may suffice in some situations.

Soil. The first reported field trial of the use of hyperaccumulating plants to remove metals from a soil contaminated by sludge applications has been reported (103). The results were positive, but the rates of metal uptake suggest a time scale of decades for complete cleanup. Trials with higher biomass plants, such as *B. juncea*, are underway at several chromium and lead contaminated sites (88), but data are not yet available.

The bacterial reduction of Cr(VI) to Cr(III) discussed above is also being used to reduce the hazards of chromium in soils and water (104).

Conclusions

Bioremediation has been successfully used to treat a wide range of contaminants, including crude oils and refined petroleum products, halogenated solvents, pesticides, herbicides, military chemicals, and mine waters. Much of this success has come by small adjustments of the local environment to encourage the growth of remediating organisms. Fertilizer addition has been a successful treatment for terrestrial and marine oil spills, and the addition of co-substrates, particularly methane, has been a successful treatment for remediating halogenated solvents, such as trichloroethylene. Composting is proving to be a successful treatment for a range of contaminants, and constructed wetlands are successfully treating a range of wastewaters, including those emanating from mines.

One area where there is considerable disagreement between academic scientists and engineering practitioners of bioremediation is the area of bioaugmentation, the addition of selected microorganisms to a site to encourage biodegradation. Dozens of companies sell bacteria for this purpose and claim success in the field, but efforts to demonstrate effectiveness rigorously have met with little success (105,57). Perhaps the most startling test was performed by the U.S. EPA when testing potential inoculants for stimulating oil biodegradation in Alaska following the *Exxon Valdez* oil spill (106). Eight products were tested in small laboratory reactors that allowed substantial degradation of a test oil by the indigenous organisms of Prince William Sound. All eight microbial inocula had a greater stimulatory effect on alkane degradation, at least for the first 11 days, when they were sterilized by autoclaving prior to addition! This suggests that the indigenous organisms readily out-competed the added products, but that autoclaving the products released some trace nutrient that was able to stimulate the growth of the endogenous organisms.

Of course, bioaugmentation may prove more effective with contaminants that have only recently entered the biosphere, such as methyl-*tert*-butyl ether, or with organisms that have a dramatic selective advantage over indigenous organisms. For example, modern molecular biology may offer opportunities for moving effective degradation pathways into organisms native to a contaminated site, improving biodegradation pathways by broadening the substrate range of degradative enzymes, or removing regulatory constraints to maximize degradative activities. There are, however, many technical and regulatory hurdles to be surmounted before this potential can be tested.

Another area where there is controversy is in the role of surfactants. Many bacteria, particularly those that degrade hydrocarbons, produce surfactants as they grow. Some release them into the medium, others incorporate them into their cell exterior, and there have been elegant experiments to show that inhibiting the production of these compounds inhibits the ability of the bacteria to degrade oil. There have thus been many suggestions to add surfactants, either bacterial or synthetic, to stimulate biodegradation. This seems to be beneficial in the case of some oil dispersants at sea (12), but there have not yet been any clear demonstrations of efficacy on a large scale in soils. The role of surfactants in bioremediation is an area that requires further study.

Bioremediation has many advantages over other technologies, both in cost and in effectively destroying or extracting the pollutant. An important issue is thus when to consider it, and a series of questions may lead to the appropriate answer (see Table 4).

Table 4. Will Bioremediation be a Suitable Treatment for a Site Contaminated with Organic, Nitrogenous, or Organic Contaminants?

<i>Organic</i>	
Is the contaminant biodegradable? If the contaminant is a complex mixture of components, are the individual chemical species biodegradable? If the contaminant has been at the site for some time, biodegradation of the most readily degradable components may have already occurred. Is the residual contamination biodegradable?	
Are degrading organisms present at the site?	
What is limiting their growth and activity? Can this be added effectively?	
Are the levels of contaminant amenable to bioremediation? Are they toxic to microorganisms? Are they so abundant that even substantial microbial activity will take too long to clean the site?	
Are the clean-up standards reasonable? Are biological processes known to degrade substrates down to the levels required?	
<i>Nitrogenous</i>	
Are appropriate microorganisms present at the site?	
What is limiting their growth and activity? Can this be added effectively?	
Are the levels of contaminant amenable to bioremediation? Are they toxic to microorganisms? Are they so abundant that even substantial microbial activity will take too long to clean the site?	
Can the nitrogenous compound be used by plants?	
Are the clean-up standards reasonable? Are biological processes known to degrade substrates down to the levels required?	
<i>Inorganic</i>	
Can the contaminant be made less hazardous by changing its redox state?	
Can the contaminant be brought to a reactor or constructed wetland where biological systems, microbial or plant, can extract and immobilize the contaminant?	
Can plants extract the contaminant from the soil matrix?	
Are the clean-up standards reasonable? Are biological processes known to accumulate contaminants down to the levels required?	

If the answers to the questions in Table 4 lead to the selection of bioremediation, it then becomes important to assess the success of the bioremediation strategy in achieving the clean-up criteria. A major disadvantage of bioremediation is that it is typically rather slower than competing technologies such as thermal treatments. How can regulators and responsible parties gain confidence during this time that success will indeed be achieved? The National Research Council (107) has recently addressed this issue, and suggested a three-fold strategy for "proving" bioremediation: (1) a documented loss of contaminants from the site; (2) laboratory tests showing the potential of endogenous microbes to catalyze the reactions of interest; and (3) some evidence that this potential is achieved in the field.

Although laboratory tests to demonstrate the potential of endogenous organisms are relatively straightforward, the other two are often not as simple as they seem, especially in soils and sediments. Documenting the loss of contaminant is often difficult because of the heterogeneous distribution of the contaminant in the field; large numbers of samples may be needed in order to be able to detect statistically significant decreases in absolute concentration. Providing evidence that biodegradation has indeed been stimulated is also a challenge, but if the contaminant is a complex mixture, such as a crude oil, refined petroleum, or a mixture of polychlorinated biphenyls, the least degradable compounds in the mixture can be used as conserved internal markers for

quantifying the degradation of the more degradable components.

For example, hopane (Fig. 12) is a conserved marker in crude oils in at least the early stages of biodegradation (up to 80% degradation) (108); its concentration increases in the residual oil as biodegradation proceeds. Basing estimates of biodegradation on hopane allowed us to quantify the effect of the successful bioremediation strategy following the *Exxon Valdez* oil-spill (14,15). Even in refined products, the least degradable detectable analyte can serve this role. Thus, the trimethyl-phenanthrenes can be used to estimate qualitatively the degradation of diesel oils in the environment, even though these molecules are themselves biodegradable (109). Of course, since these molecules are themselves biodegradable, estimates of the rates of biodegradation of more readily biodegradable compounds are systematically underestimated using this approach, but it can still provide very valuable information. The more halogenated polychlorinated biphenyls can serve a similar role in assessing the environmental fate of these products (55), and benzene is typically the last compound in BTEX plumes to be degraded, particularly under anaerobic conditions.

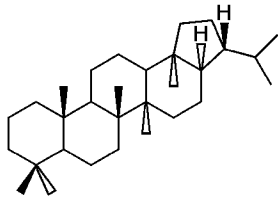


Fig. 12. 17α(H),21β(H)-hopane.

In situations where conserved internal markers cannot be used, such as in spills of essentially pure compounds, the evidence for enhanced biodegradation may have to be more indirect. Oxygen consumption, increases in microbial activity or population, and carbon dioxide evolution have all been used with success.

Finally, a caveat. Despite its documented success in many situations, bioremediation may not always be able to meet current clean-up criteria for a particular site. Some standards are so tight that they are essentially "detection limit" standards, and it is not clear that biological processes will be able to remove contaminants to such low levels. For example, the level of contaminant may be so low that it does not induce the microorganisms to produce the enzymes necessary for biodegradation. Or perhaps the contaminant is bound to soil or sediment particles in such a way that it is not available for biodegradation, although it is still extractable with aggressive solvents in analytical procedures. These are areas that require further research, but bioremediation will be more likely to fulfill its promise as an important tool in contaminated site remediation if there is progress towards standards based on bioavailability and net environmental benefit from the clean up, rather than on arbitrary absolute standards.

BIBLIOGRAPHY

1. K. Devine, in R. E. Hinchee, J. A. Kittel, and H. J. Reisinger, eds., *Applied Bioremediation of Hydrocarbons*, Battelle Press, Columbus, Ohio, 1995, pp. 53–59.
2. S. S. Butcher, R. J. Charlson, G. H. Orians, and G. V. Wolfe, *Global Biogeochemical Cycles*, Academic Press, New York, 1992.
3. I. D. Prijambada, S. Negoro, T. Yomo, and I. Urabe, *Appl. Environ. Microbiol.* **61**, 2020–2022 (1995).
4. P. Adriaens and T. M. Vogel, in L. Y. Young and C. E. Cerniglia, eds., *Microbial Transformation and Degradation of Toxic Organics*, Wiley-Liss, New York, 1995, pp. 435–486.
5. A. Paszczynski and R. L. Crawford, *Biotechnol. Progr.* **11**, 368–379 (1995).
6. J. J. Childress, *Amer. Zool.* **35**, 83–90 (1995).
7. T. D. Sharkey and E. L. Singsaas, *Nature* **374**, 769 (1995).
8. R. C. Prince, Bioremediation of Crude Oil, in R. A. Meyers, ed., *Encyclopedia of Environmental Analysis and Remediation*, John Wiley & Sons, Inc. (in press) (1998).
9. J. P. Salanitro, *Curr. Opin. Biotechnol.* **6**, 337–340 (1995).
10. J. M. Yudelson and P. D. Tinari, in R. E. Hinchee, G. D. Sayles and R. S. Skeen, eds., *Biological Unit Processes for Hazardous Waste Treatment*, Battelle Press, Columbus, Ohio, 1995, pp. 205–209.
11. National Research Council *Oil in the Sea: Inputs, Fates and Effects*, National Academy Press, Washington, D.C., 1985.
12. R. Varadaraj, M. L. Robbins, J. Bock, S. Pace, and D. MacDonald, in *Proceedings of the 1995 International Oil Spill Conference*, American Petroleum Institute, Washington, D.C., 1995, pp. 101–106.
13. R. C. Prince, *Critical Reviews Microbiology* **19**, 217–242 (1993).
14. R. C. Prince, J. R. Clark, J. E. Lindstrom, E. L. Butler, E. J. Brown, G. Winter, M. J. Grossman, R. R. Parrish, R. E. Bare, J. F. Braddock, W. G. Steinhauer, G. S. Douglas, J. M. Kennedy, P. Barter, J. R. Bragg, E. J. Harner, and R. M. Atlas, in R. E. Hinchee, B. C. Alleman, R. E. Hoeppel and R. N. Miller, eds., *Hydrocarbon Remediation*, Lewis Publishers, Boca Raton, Fla., 1994, pp. 107–124.
15. J. R. Bragg, R. C. Prince, E. J. Harner, and R. M. Atlas, *Nature* **368**, 413–418 (1994).
16. R. C. Borden, C. A. Gomez, and M. T. Becker, *Ground Water* **33**, 180–189 (1995).
17. R. D. Norris, in Ref. 1, pp. 483–487 (1995).
18. G. Battermann and M. Meier-Luhr, in Ref. 1, pp. 155–164.
19. D. R. Lovley, J. C. Woodward, and F. H. Chapelle, *Nature* **370**, 128–131 (1994).
20. B. A. Keet, in Ref. 1, pp. 329–334.
21. S. J. McMillen, N. R. Gray, J. M. Kerr, A. G. Requejo, T. J. McDonald, and G. S. Douglas, in R. E. Hinchee, G. S. Douglas, and S. K. Ong, eds., *Monitoring and Verification of Bioremediation*, Battelle Press, Columbus, Ohio, 1995, pp. 1–9.
22. B. M. Peyton, M. J. Trues, R. S. Skeen, and B. S. Hooker, in R. E. Hinchee, A. Leeson, and L. Semprini, eds., *Bioremediation of Chlorinated Solvents*, Battelle Press, Columbus, Ohio, 1995, pp. 111–116.
23. D. Kohler-Staub, S. Frank, and T. Leisinger, *Biodegradation* **6**, 229–236 (1995).
24. D. B. Janssen, F. Pries, and J. R. van der Ploeg, *Ann. Rev. Microbiol.* **48**, 163–191 (1994).
25. D. J. Arp, *Curr. Opin. Biotechnol.* **6**, 352–358 (1995).
26. S. J. Ergas, K. Kinney, M. E. Fuller, and K. M. Scow, *Biotechnol. Bioengineer.* **44**, 1048–1054 (1994).
27. K. H. Lombard, J. W. Borten, and T. C. Hazen, in R. E. Hinchee, ed., *Air Sparging for Site Remediation*, Lewis Publishers, Boca Raton, Fla., 1994, pp. 81–96.
28. T. A. Anderson and B. T. Walton, *Environ. Toxicol. Chem.* **14**, 2041–2047 (1995).
29. J. A. Oleskiewicz and M. Elektorowicz, *J. Soil Contam.* **2**, 205–208 (1993).
30. W. M. Wu, J. Nye, R. F. Hickey, M. K. Jain and J. G. Zeikus, in R. E. Hinchee, A. Leeson and L. Semprini, eds., *Bioremediation of Chlorinated Solvents*, Battelle Press, Columbus, Ohio, 1995, pp. 45–52.
31. L. M. F. Dossantos and A. G. Livingston, *Appl. Microbiol. Biotechnol.* **42**, 421–431 (1994).

32.

M. Knippschild and H. J. Rehm, *Appl. Microbiol. Biotechnol.* **44**, 253–258 (1995).

33.

J. Gerritse, V. Renard, J. Visser, and J. C. Gottschal, *Appl. Microbiol. Biotechnol.* **43**, 920–928 (1995).

34.

G. W. Gribble, *J. Chem. Ed.* **71**, 907–911 (1994).

35.

S. Fetzner and F. Lingens, *Microbiol. Rev.* **58**, 641–673 (1994).

36.

P. Juteau, R. Beaudet, G. Mcsween, F. Lepine, and J. G. Bisaillon, *Can. J. Microbiol.* **41**, 862–868 (1995).

37.

C. Kennes, W. M. Wu, L. Bhatnagar and J. G. Zeikus, *Appl. Microbiol. Biotechnol.* **44**, 801–806 (1996).

38.

K. A. McAllister, H. Lee, and J. T. Trevors, *Biodegradation* **7**, 1–40 (1996).

39.

B. C. Okeke, A. Paterson, J. E. Smith, and I. A. Watson-Craik, *Letts. Appl. Microbiol.* **19**, 284–287 (1994).

40.

J. Aislabie and G. Lloyd-Jones, *Aust. J. Soil. Res.* **33**, 925–942 (1995).

41.

D. A. Abramowicz, *Crit. Rev. Biotechnol.* **10**, 241–251 (1990).

42.

S. Han, L. D. Eltis, K. N. Timmis, S. W. Muchmore, and J. T. Bolin, *Science* **270**, 976–980 (1995).

43.

B. N. Aronstein and L. E. Rice, *J. Chem. Technol. Biotechnol.* **63**, 321–328 (1995).

44.

J. L. Gentry and T. J. Simpkin, in R. E. Hinchee, R. M. Miller, and P. C. Johnson, eds., *In situ Aeration: Air Sparging, Bioventing and Related Remediation Processes*, Battelle Press, Columbus, Ohio, pp. 283–289 (1995).

45.

A. M. Ferro, R. C. Sims, and B. Bugbee, *J. Environ. Qual.* **23**, 272–279 (1994).

46.

S. G. Huling, D. F. Pope, J. E. Matthews, J. L. Sims, R. C. Sims, and D. L. Sorenson, in R. E. Hinchee, R. E. Hoeppeel and D. B. Anderson, eds., *Bioremediation of Recalcitrant Organics*, Battelle Press, Columbus, Ohio, 1995, pp. 101–109.

47.

G. M. Colores, P. M. Radehaus, and S. K. Schmidt, *Appl. Biochem. Biotechnol.* **54**, 271–275 (1995).

48.

M. P. Otte, J. Gagnon, Y. Comeau, N. Maatte, C. W. Greer, and R. Samson, *Appl. Microbiol. Biotechnol.* **40**, 926–932 (1994).

49.

A. J. Wall and G. W. Stratton, *Water Air Soil Poll.* **82**, 723–737 (1995).

50.

F. Baud-Grasset and T. M. Vogel, in R. E. Hinchee, J. E. Fredrickson and B. C. Alleman, eds., *Bioaugmentation for Site Remediation*, Battelle Press, Columbus, Ohio, 1995, pp. 39–48.

51.

F. C. Michel, C. A. Reddy, and L. J. Forney, *Appl. Environ. Microbiol.* **61**, 2566–2571 (1995).

52.

C. J. Gantzer and D. Cosgriff, in Ref. 44, pp. 543–549.

53.

R. U. Edgehill, *Water Res.* **30**, 357–363 (1996).

54.

D. Dietrich, W. J. Hickey, and R. Lamar, *Appl. Environ. Microbiol.* **61**, 3904–3909 (1995).

55.

D. L. Bedard and R. J. May, *Environ. Sci. Technol.* **30**, 237–245 (1996).

56.

D. L. Bedard, S. C. Bunnell and L. A. Smullen, *Environ. Sci. Technol.* **30**, 687–694 (1996).

57.

M. R. Harkness, J. B. McDermott, D. A. Abramowicz, J. J. Salvo, W. P. Flanagan, M. L. Stephens, F. J. Mondello, R. J. May, H. J. Lobos, K. M. Carroll, M. J. Brennan, A. A. Bracco, K. M. Fish, G. L. Warner, P. R. Wilson, D. K. Dietrich, D. T. Lin, C. B. Morgan, C. B. and W. L. Gately, *Science* **259**, 503–507 (1993).

58.

R. E. Dick and J. P. Quinn, *Appl. Microbiol. Biotechnol.* **43**, 545–550 (1995).

59.

R. H. Kaake, D. L. Crawford, and R. L. Crawford, *Biodegradation* **6**, 329–337 (1995).

60.

T. B. Hammill and R. L. Crawford, *Appl. Environ. Microbiol.* **62**, 1842–1846 (1996).

61.

J. Kazumi and D. G. Capone, *Appl. Environ. Microbiol.* **62**, 2820–2829 (1995).

62.

I. Mirgain, G. Green, and H. Monteil, *Environ. Technol.* **16**, 967–976 (1995).

63.

C. J. Hapeman, J. S. Kams, and D. R. Shelton, *J. Agric. Food Chem.* **43**, 1383–1391 (1995).

64.

D. R. Shelton, S. Khader, J. S. Kams, and B. M. Pogell, *Biodegradation* **7**, 129–136 (1996).

65.

J. A. Entry, P. K. Donnelly, and W. H. Emmingham, *Appl. Soil Ecol.* **3**, 85–90 (1996).

66.

D. L. Kaplan, *Curr. Opinion Biotechnol.* **3**, 253–260 (1992).

67.

T. Gorontzy, O. Dryzga, M. W. Kahl, D. Bruns-Nagel, J. Breitung, E. von Loew and K.-H. Blotevogel, *Crit. Rev. Microbiol.* **20**, 265–284 (1994).

68.

N. S. Kasimov, V. B. Grebenyuk, T. V. Koroleva, and Y. V. Proskuryakov, *Eurasian Soil Science* **28**, 79–95 (1996).

69.

U.S. Pat. 5,387,271 (1995), D. L. Crawford, T. O. Stevens and R. L. Crawford.

70.

S. Eversmeyer, P. Faessler, and C. Bird, *Soil Groundwater Cleanup* (Dec. 27–29, 1995).

71.

E. Flathman, B. J. Krupp, P. Zottola, J. R. Trausch, J. H. Carson, R. Yao, G. J. Laird, P. M. Woodhull, D. E. Jerger, and P. R. Lear, *Remediation* **6**, 57–79 (1996).

72.

Z. Ronen, J. M. Bollag, C. H. Hsu, and J. C. Young, *Ground Water* **34**, 194–199 (1996).

73.

M. H. Essa, S. Farooq, and G. F. Nakhla, *Water Air Soil Poll.* **87**, 267–281 (1996).

74.

V. Cuomo, A. Perretti, I. Palomba, A. Verde, and A. Cuomo, *J. Appl. Phycol.* **7**, 479–485 (1995).

75.

G. C. Schmidt and M. B. Ballew, in R. E. Hinchee, J. L. Means, and D. R. Burris, eds., *Bioremediation of Inorganics*, Battelle Press, Columbus, Ohio, 1995, pp. 109–116.

76.

W. J. Hunter and R. F. Folett, in *Clean Water—Clean Environment—21st Century*, Vol. II: *Nutrients*. American Society of Agricultural Engineers, St Joseph, Mich., 1995, pp. 79–82.

77.

D. K. Ewart and M. N. Hughes, *Adv. Inorg. Chem.* **36**, 103–135 (1991).

78.

B. Volesky and Z. R. Holan, *Biotechnol. Prog.* **11**, 235–250 (1995).

79.

C. E. Turick, W. A. Apel, and N. S. Carmiol, *Appl. Microbiol. Biotechnol.* **44**, 683–688 (1996).

80.

W. T. Frankenberger and U. Karlson, *Geomicrobiol. J.* **12**, 265–278 (1994).

81.

D. Ahmann, A. L. Roberts, L. R. Knumholz, and F. M. M. Morel, *Nature* **371**, 750 (1994).

82.

K. A. Rittle, J. I. Drever, and P. J. S. Colberg, *Geomicrobiol. J.* **13**, 1–11 (1995).

83.

L. E. Macaskie, R. M. Empson, F. Lin, and M. R. Tolley, *J. Chem. Technol. Biotechnol.* **63**, 1–16 (1995).

84.

K. M. Bonthron, G. Basnakova, F. Lin, and L. E. Macaskie, *Nature Biotechnology* **14**, 635–638 (1996).

85.

E. J. P. Phillips, E. R. Landa, and D. R. Lovley, *J. Ind. Microbiol.* **14**, 203–207 (1995).

86.

E. Saouter, M. Gillman, and T. Barkay, *J. Ind. Microbiol.* **14**, 343–348 (1995).

87.

V. Dushenkov, P. B. A. N. Kumar, H. Motto, and I. Raskin, *Environ. Sci. Technol.* **29**, 1239–1245 (1995).

88.

D. E. Salt, M. Blaylock, N. P. B. A. Kumar, V. Dushenkov, B. D. Ensley, I. Chet, and I. Raskin, *Bio-Technology* **13**, 468–474 (1995).

89.

W. E. Rauser, *Ann. Rev. Biochem.* **59**, 61–86 (1990).

90.

D. E. Salt, R. C. Prince, I. J. Pickering, and I. Raskin, *Plant Physiol.* **109**, 1427–1433 (1995).

91.

U. Kramer, J. D. Cotterhowells, J. M. Chamock, A. J. M. Baker, and J. A. C. Smith, *Nature* **379**, 635–638 (1996).

92.

G. S. Banuelos and D. W. Meek, *J. Environ. Qual.* **19**, 772–777 (1990).

93.

A. M. Zayed and N. Terry, *J. Plant. Physiol.* **143**, 8–14 (1994).

94.

C. L. Rugh, H. D. Wilde, N. M. Stack, D. M. Thomson, A. O. Summers, and R. B. Meagher, *Proc. Natl. Acad. Sci. USA* **93**, 3182–3187 (1996).

95.

L. J. Barnes, P. J. M. Scheeren, and C. J. N. Buisman, in J. L. Means and R. E. Hinchee, eds., *Emerging Technology for Bioremediation of Metals*, Lewis Publishers, Boca Raton, Fla., 1994, pp. 38–49.

96.

J. M. Macy, S. Lawson, and H. DeMott-Decker, *Appl. Microbiol. Biotechnol.* **40**, 588–594 (1993).

97.

L. P. Owens, K. C. Kovac, J. A. L. Kipps, and D. W. J. Hayes, in Ref. 75, pp. 89–94.

98. J. L. Kipps, in J. L. Means and R. E. Hinchee, eds., *Emerging Technology for Bioremediation of Metals*, Lewis Publishers, Boca Raton, Fla., 1994, pp. 105–109.
99. C. M. Cooney, *Environ. Sci. Technol.* **30**, A194 (1996).
100. G. Robb and J. Robinson, *Mining Environ. Manage.*, 19–21 (Sept. 1995).
101. M. E. Dodds-Smith, C. A. Payne, and J. J. Gusek, *Mining Environ. Manage.*, 22–24 (Sept. 1995).
102. J. Bender and P. Phillips, *Mining Environ. Manage.*, 25–27 (Sept. 1995).
103. A. J. M. Baker, S. P. McGrath, C. M. D. Sidoli, and R. D. Reeves, *Resource Conserv. Recycling* **11**, 41–49 (1994).
104. L. J. DeFilippi, *Environ. Sci. Pollut. Control Ser.* **8**, 437–457 (1994).
105. P. H. Pritchard, *Curr. Opinion Biotechnol.* **3**, 232–243 (1992).
106. A. D. Venosa, J. R. Haines, W. Nisamaneepong, R. Govind, S. Pradhan, and B. Siddique, *J. Ind. Microbiol.* **10**, 13–23 (1992).
107. National Research Council, *In situ Bioremediation. When Does it Work?* National Academy Press, Washington, D.C., 1993.
108. R. C. Prince, D. L. Elmendorf, J. R. Lute, C. S. Hsu, C. E. Haith, J. D. Senius, G. J. Dechert, G. S. Douglas, and E. L. Butler, *Environ. Sci. Technol.* **28**, 142–145 (1994).
109. G. S. Douglas, K. J. McCarthy, D. T. Dahlen, J. A. Seavey, W. G. Steinhauer, R. C. Prince, and D. L. Elmendorf, *J. Soil Contam.* **1**, 197–216 (1992).

General References

M. Alexander, *Biodegradation and Bioremediation*, Academic Press, San Diego, Calif., 1994.

R. E. Hinchee, J. T. Wilson and D. C. Downey, eds., *Intrinsic Bioremediation*, Battelle Press, Columbus, Ohio, 1995.

R. E. Hinchee, R. M. Miller, and P. C. Johnson, eds., *In situ Aeration: Air Sparging, Bioventing and Related Remediation Processes*, Battelle Press, Columbus, Ohio, 1995.

R. E. Hinchee, J. E. Fredrickson, and B. C. Alleman, eds., *Bioaugmentation for Site Remediation*, Battelle Press, Columbus, Ohio, 1995.

R. E. Hinchee, A. Leeson, and L. Semprini, eds., *Bioremediation of Chlorinated Solvents*, Battelle Press, Columbus, Ohio, 1995.

R. E. Hinchee, G. S. Douglas, and S. K. Ong, eds., *Monitoring and Verification of Bioremediation*, Battelle Press, Columbus, Ohio, 1995.

R. E. Hinchee, J. A. Kittel, and H. J. Reisinger, eds., *Applied Bioremediation of Hydrocarbons*, Battelle Press, Columbus, Ohio, 1995.

R. E. Hinchee, R. E. Hoeppeel, and D. B. Anderson, eds., *Bioremediation of Recalcitrant Organics*, Battelle Press, Columbus, Ohio, 1995.

R. E. Hinchee, C. M. Vogel, and F. J. Brockman, eds., *Microbial Processes for Bioremediation*, Battelle Press, Columbus, Ohio, 1995.

R. E. Hinchee, G. D. Sayles, and R. S. Keen, eds., *Biological Unit Processes for Hazardous Waste Treatment*, Battelle Press, Columbus, Ohio, 1995.

R. E. Hinchee, J. L. Means, and D. R. Burris, eds., *Bioremediation of Inorganics*, Battelle Press, Columbus, Ohio, 1995.

National Research Council, *In situ Bioremediation. When Does it Work?* National Academy Press, Washington, D.C., 1993.

Organization for Economic Cooperation and Development, *Bioremediation: the Tokyo '94 Workshop*, OECD, Paris, 1995.

L. Y. Young and C. E. Cerniglia, eds., *Microbial Transformation and Degradation of Toxic Organic Chemicals*, Wiley-Liss, New York, 1995.

R. E. Hinchee, A. Leeson, L. Semprini, and S. K. Ong, eds., *Bioremediation of Chlorinated and Polycyclic Aromatic Hydrocarbons*, Lewis Publishers, Ann Arbor, Mich., 1994.

R. E. Hinchee, B. C. Alleman, R. E. Hoeppeel, and R. N. Miller, eds., *Hydrocarbon Bioremediation*, Lewis Publishers, Ann Arbor, Mich., 1994.

R. E. Hinchee, D. B. Anderson, F. B. Metting, Jr., and G. D. Sayles, eds., *Applied Biotechnology for Site Remediation*, Lewis Publishers, Ann Arbor, Mich., 1994.

J. L. Means and R. E. Hinchee, eds., *Emerging Technology for Bioremediation of Metals*, Lewis Publishers, Ann Arbor, Mich., 1994.

R. E. Hinchee, ed., *Air Sparging for Site Remediation*, Lewis Publishers, Ann Arbor, Mich., 1994.

R. F. Hickey and G. Smith, eds., *Biotechnology in Industrial Waste Treatment and Bioremediation*, Lewis Publishers, Boca Raton, Fla., 1996.

R. H. Kadlec and R. L. Knight, *Treatment Wetlands*, Lewis Publishers, Boca Raton, Fla., 1996.

Roger C. Prince
 Exxon Research and Engineering Company

BIOSEPARATIONS

The large-scale purification of proteins and other bioproducts is the final production step, prior to product packaging, in the manufacture of therapeutic proteins, specialty enzymes, diagnostic products, and value-added products from agriculture. These separation steps, taken to purify biological molecules or compounds obtained from biological sources, are referred to as bioseparations. Large-scale bioseparations combine art and science. Bioseparations often evolve from laboratory-scale techniques, adapted and scaled up to satisfy the need for larger amounts of extremely pure test quantities of the product. Uncompromising standards for product quality, driven by commercial competition, applications, and regulatory oversight, provide many challenges to the scale-up of protein purification. The rigorous quality control embodied in current good manufacturing practices, and the complexity and lability of the macromolecules being processed provide other practical issues to address (1).

Recovery and purification of new biotechnology products is the fastest growing area of bioseparations. Biotechnology was broadly defined in 1991 by the U.S. Office of Technology Assessment as "any technique that uses living organisms (or parts of organisms) to make or modify products, to improve plants or animals, or to develop microorganisms for specific uses." The new biotechnology, introduced in 1970, involves directed manipulation of the cell's genetic machinery through recombinant deoxyribonucleic acid (DNA) techniques and cell fusion. The new biotechnology was first applied on an industrial scale in 1979. Since then it has fundamentally expanded the utility of biological systems, so that biological molecules for which there is no other means of industrial production can be generated. Substantial manufacturing capability is expected to be needed to bring about the full application of this biotechnology (2). The recovery, purification, and packaging of biotechnological products for delivery to the consumer is undergoing unprecedented growth.

Manufacturing approaches for selected bioproducts of the new biotechnology impact product recovery and purification. The most prevalent bioseparations method is chromatography (qv). Thus the practical tools used to initiate scaleup of process liquid chromatographic separations starting from a minimum amount of laboratory data are given.

Economic Aspects

The development of biotechnology processes in the biopharmaceutical and bioproduct industries is driven by the precept of being first to market while achieving a defined product purity, and developing a reliable process to meet validation requirements. The economics of bioseparations are important, but are likely to be secondary to the goal of being first to market. The cost of a lost opportunity in a tightly focused market where there is room for only a few manufacturers can be devastating for products which take 5 to 10 years and \$100 to \$200 × 10⁶ to develop. After process and product are validated, the cost of change in any portion of the procedure can also be great, if only to satisfy regulatory constraints. Hence, once the manufacturing process is in place, changes are likely to be considered only if significant improvements result.

The three main sources of competitive advantage in the manufacture of high value protein products are first to market, high product quality, and low cost (3). The first company to market a new protein biopharmaceutical, and the first to gain patent protection, enjoys a substantial advantage. The second company to enter the market may find itself enjoying only one-tenth of the sales. In the absence of patent protection, product differentiation becomes very important. Differentiation reflects a product that is purer, more active, or has a greater lot-to-lot consistency.

Biopharmaceuticals and Protein Products. Purification of proteins is a critical and expensive part of the production process, often accounting for >50% of total production costs (2). Hence, bioseparation processes have a significant impact on manufacturing costs. For small-volume, very high value biotherapeutics (Table 1), however, these costs may be considered secondary to the first to market principle unless a lower cost competitor surfaces. Annual 1995 sales were \$700 million for human insulin (5), \$300 million for tissue plasminogen activator, and \$220 million for human growth hormone (6). The most successful bioproduct in biotechnology history, recombinant erythropoetin (EPO), had worldwide sales estimated at \$1.6 to \$2.6 billion in 1995 (5,7,8). Epogen is a genetically engineered version of erythropoetin [11096-26-7], which is produced by the kidneys and stimulates blood stem cells to mature into red blood cells. Epogen can reverse the severe anemia often caused by kidney disease. Amgen's sales of this product, together with Neupogen (a recombinant protein that directs blood stem cells to become bacteria-fighting neutrophils), was about \$1.8 billion in 1995 (9).

Table 1. Unit Values and Relative Production Quantities for Selected Approved Biopharmaceuticals, 1990–1991^a

Product	Year approved	Selling price, \$ g ^b	Quantity for \$200 × 10 ⁶ in sales, kg
human insulin	1982	375	530.0
tissue plasminogen activator	1987	23,000	8.7
human growth hormone	1985	35,000	5.7
erythropoetin (Epogen)	1989	840,000	0.24
GM CSF	1991	384,000	0.52
G-CSF	1991	450,000	0.44

^a Adapted from Ref. 2 with additional data from Ref. 4.

^b Values are approximate and are likely to decrease.

Bioproduct Separations

The task of quickly specifying, designing, and scaling-up a bioproduct separation is not simple. These separations are carried out in a liquid phase using macromolecules which are labile, and where conformation and heterogeneous chemical structure undergoing even subtle change during purification may result in an unacceptable product. A typical purification scheme for biopharmaceutical proteins involves the harvesting of protein-containing material or cells, concentration of protein using ultrafiltration (qv), initial chromatographic steps, viral clearance steps, additional chromatographic steps, again concentration of protein using ultrafiltration, and finally formulation (10).

Biosynthetic Human Insulin from *E. coli*. Insulin [9004-10-8], a polypeptide hormone, stimulates anabolic reactions for carbohydrates, proteins, and fats thereby producing a lowered blood glucose level. Porcine insulin [12584-58-6] and bovine insulin [11070-73-8] were used to treat diabetes prior to the availability of human insulin [11061-68-0]. All three insulins are similar in amino acid sequence. Eli Lilly's human insulin was approved for testing in humans in 1980 by the U.S. FDA and was placed on the market by 1982 (11,12).

Human insulin was the first animal protein to be made in bacteria in a sequence identical to the human pancreatic peptide. Expression of separate insulin A and B chains were achieved in *Escherichia coli* K-12 using genes for the insulin A and B chains synthesized and cloned in frame with the

β -galactosidase gene of plasmid pBR322 (13,14). Insulin's small size, 21 amino acids for the A-chain, mol wt 2300; and 30 for the B-chain, mol wt 3400, together with the absence of methionine (Met) and tryptophan (Trp) residues, were critical elements both in the decision to undertake cloning of this peptide hormone and in the rapid development of the manufacturing process. The Met and Trp residues, produced as a consequence of engineering and expression in *E. coli*, are hydrolyzed by reagents used during the recovery process. The presence of these amino acids in insulin would have resulted in the hydrolysis and destruction of the product (12).

Recovery and Purification. The production of Eli Lilly's human insulin requires 31 principal processing steps of which 27 are associated with product recovery and purification (13). The production process for human insulin, based on a fermentation which yields proinsulin, provides an instructive case study on the range of unit operations which must be considered in the recovery and purification of a recombinant product from a bacterial fermentation. Whereas the exact sequence has not been published, the principle steps in the purification scheme are outlined in Figure 1a.

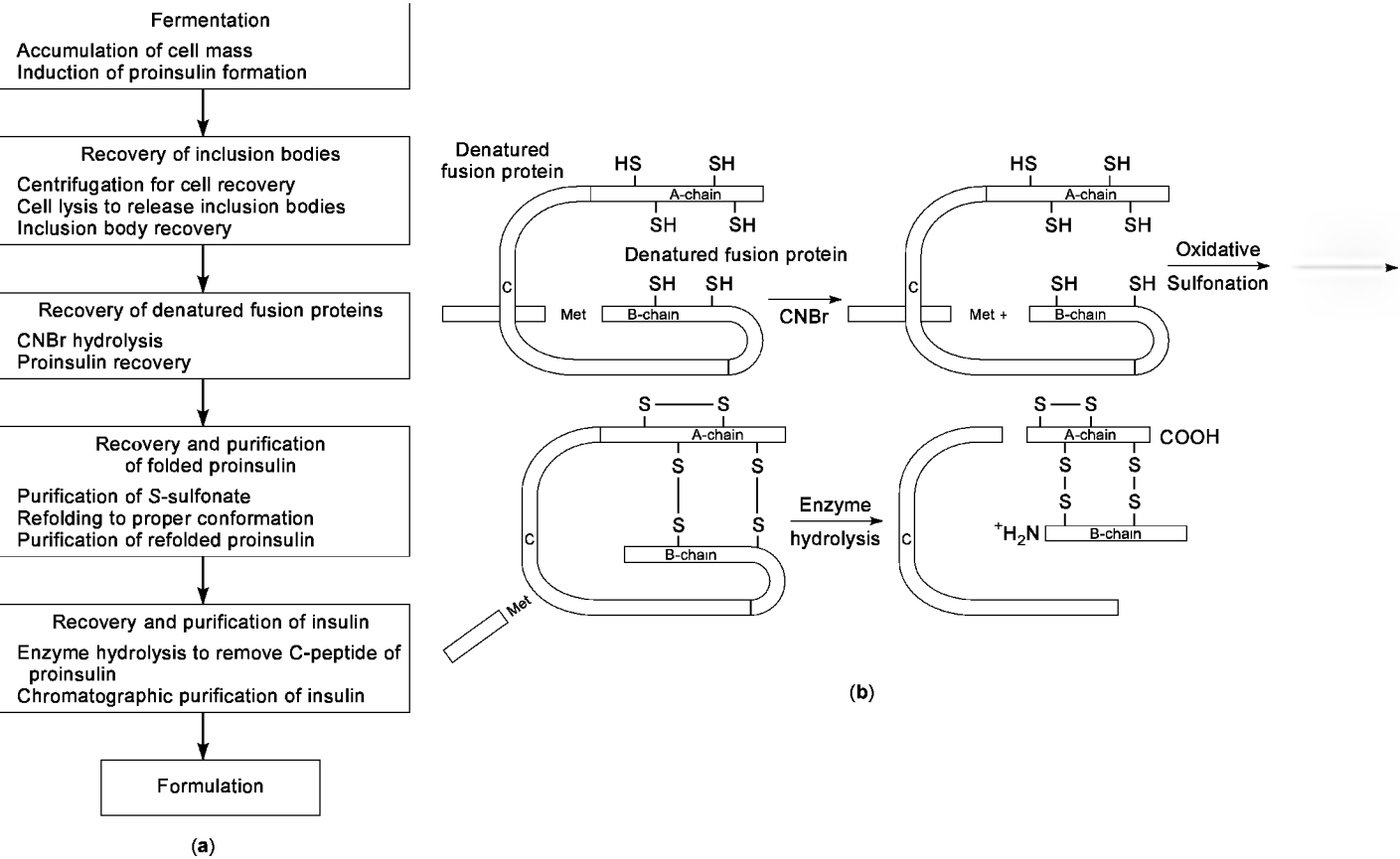


Fig. 1. (a) Process flow sheet for human insulin production, recovery, and purification (12); (b) corresponding steps in recovery of biosynthetic human insulin.

The fermentation product is a fusion protein where a portion of the Trp protein is connected to proinsulin through a Met residue (Fig. 1b). The *E. coli* contains a plasmid having the proinsulin gene connected to the Trp promoter. The Trp operon is turned on when the fermentation media is allowed to become depleted of tryptophan and the production of a fusion protein of proinsulin occurs. An inclusion body, ie, a large body of aggregated protein and nucleic acids occupying about half of the cell volume is formed. Because formation of inclusion bodies causes cell growth to stop, premature formation of inclusion bodies results in lower productivity. Hence, the Trp switch is an important practical tool in maximizing productivity. At this point the fermentation is complete, and protein recovery, dissolution, protein refolding, and purification is carried out (12). Following inclusion body recovery, CNBr, a hydrolytic agent which specifically attacks Met and Trp linkages, cleaves away the fusion protein from the proinsulin (see Fig. 1b). No Met or Trp occurs in proinsulin, so the proinsulin molecule is left intact. The proinsulin is then subjected to oxidative sulfitylase, refolding to its proper conformation, purification, and enzyme treatment to remove the peptide connecting the insulin A- and B-chains. The crude insulin consisting of A- and B-chains in their proper conformation is then further purified using a sequence of ion-exchange (qv), reversed-phase, and size-exclusion chromatography (3,12,14).

Desamidation of asparagine or glutamine residues can occur readily in either acidic or neutral solutions; disulfide exchange reactions can occur at alkaline pH and cause formation of isomeric monomers or aggregated forms (multimers) of the protein (13). Deamidation products of insulin, referred to as desamido insulin, can form during processing. These insulin variants require high resolution chromatography techniques to remove. Therefore, a multimodal sequence of chromatographic separations for the crude recombinant insulin is required.

Ion-exchange chromatography removes most of the impurities and is followed by reversed-phase chromatography which separates insulin from structurally similar insulin-like components. Then size-exclusion (gel-permeation) chromatography is introduced to remove salts and other small molecules from the insulin. The best pH range for the acetonitrile mobile phase for reversed-phase chromatography is reported to be 3.0–4.0. This is well below the isoelectric pH of 5.4, gives excellent resolution, and minimizes deamidation of the insulin if the residue time in the reversed-phase column is less than several hours (12,14). This sequence (14) follows the principle of orthogonality of separation sequence, ie, each step is based on a different property, in this case charge, solubility, and size, respectively (1). Near the end of the chromatography sequence, the insulin may be concentrated by precipitation to form insulin zinc crystals.

Process Equipment Volumes. Product recovery involves cell lysis, centrifugation, refolding, buffer exchange, chromatography, precipitation, and filtration. Some of these steps are repeated. The volumes of the individual chromatography columns are estimated to range from 50 to 1000 L. These volumes are small compared to other types of chemical recovery processes, but are large in the context of biotechnology manufacturing. For example, the reversed-phase step uses an 80 L column volume for an insulin loading of 1.2 kg per run (3). Assuming the total amount of recombinant insulin produced annually in a typical plant is on the order of 1000 to 2000 kg, this size column would enable processing as much as 30% (>300 kg) of the annual output of insulin.

Yield Losses. The numerous steps incur a built-in yield loss. For example, if only 2% yield loss were to be associated with each step, the overall yield for a purification sequence of 10 steps would be as in equation 1:

$$\eta = 100(1 - L/100)^n = 100(1 - 0.02)^{10} = 81.7\% \quad (1)$$

where η denotes yield, L the percent yield loss, and n the number of steps. If the yield loss at each step were 5%, the overall yield would only be 60%. Maximizing recovery at each step is important.

The purification of human insulin involves five separate alterations in the molecular structure, and hence, changes in physicochemical properties during its recovery and purification. The various forms are fusion protein, denatured aggregate, denatured monomers, properly folded proinsulin, and finally insulin. Whereas various purification procedures are used repeatedly, thus introducing more steps in the process, the change of removing contaminants is maximized because the contaminants are not as likely to change chemically in the same way as the insulin molecule. The final purification steps rely on multiple properties of the insulin, such as size, hydrophobicity, ionic charge, and crystallizability (13). The final purity level is reported to be >99.99% (15).

Tissue Plasminogen Activator from Mammalian Cell Culture. Tissue-type plasminogen activator or tissue plasminogen activator [105857-23-6] (t-PA) was originally identified in tissue extracts in the late 1940s (15). Other known plasminogen activators include streptokinase from bacteria, urokinase from urine, and prourokinase from plasma (16). In 1981 the Bowes melanoma cell line was found to secrete t-PA (known as mt-PA) at 100 × higher concentrations, making possible the isolation and purification of this enzyme in sufficient quantities that antibodies could be generated and assays developed to lead to cloning of the gene for this enzyme and subsequent expression of the enzyme in both *E. coli* and a Chinese hamster ovary (CHO) cell line (15,17).

Comparison of the melanoma and recombinant forms of the enzymes showed the same activity toward dissolution of blood clots. Comparison to urokinase, another thrombolytic agent, served as the basis for introducing recombinant t-PA into clinical trials in 1984 (17). Two pilot studies demonstrated that mt-PA resulted in thrombolysis without significant fibrinogenolysis. Fibrinogen, the precursor to fibrin, is important to the clotting of blood. Because mt-PA was available in limited quantities, recombinant t-PA (rt-PA) was used to carry out the first significant clinical trial. Doses of 0.375–0.75 mg rt-PA/kg body weight was found to be effective in humans for achieving 70% recanalization. Another pilot study confirmed that a dose of 80 mg over three hours gave the same results (17). The comparison of rt-PA (injected intravenously) to streptokinase IV (injected intracoronary) produced sufficiently favorable results to end the trial early, and make the results public in 1985, resulting in the use of t-PA for heart attacks (15). A trial completed in 1996 showed that t-PA administered within three hours of a stroke caused by a clot in the brain facilitated full recovery of 31% of stroke patients. Hence, another use of rt-PA is likely to develop (18).

Approximately 500,000 Americans suffer strokes each year. Many of the 80% that survive suffer paralysis and impaired vision and speech, often needing rehabilitation and/or long-term care. Hence, whereas treatment using rt-PA is likely to be expensive (costs are \$2200/dose for treating heart attacks), the benefits of rt-PA could outweigh costs. In the case of heart attacks, the 10 times less expensive microbially derived streptokinase can be used. There is currently no competing pharmaceutical for treatment of strokes (18,19). Consequently, the cost of manufacture of rt-PA may not be as dominant an issue as would be the case of other types of bioproducts.

Characteristics of t-PA. Tissue plasminogen activator, a proteolytic, hydrophobic enzyme, has a molecular weight of 66,000, 12 disulfide bonds, four possible glycosylation sites, and a bridge of 6 amino acids connecting the principal protein structures (17,20,21). Only three of the sites (Asn-117–118, -448) are actually glycosylated (16). When administered to heart attack victims it dissolves clots consisting of platelets in a fibrin protein matrix and acts by clipping plasminogen, an active precursor protein found in the blood, to form plasmin, a potent protease that degrades fibrin (17,21). Whereas plasminogen activator is found in blood and tissues, concentrations are low (17).

The concentration of t-PA in human blood is 2–5 ng/mL, ie, 2–5 ppb. Plasminogen activation is accelerated in the presence of a clot, but the rate is slow. The dissolution of a clot requires a week or more during normal repair of vascular damage (17). Prevention of irreversible tissue damage during a heart attack requires that a clot, formed by rupture of an atherosclerotic plaque, be dissolved in a matter of hours. This rapid thrombolysis (dissolution of the clot) must be achieved without significant fibrinogenolysis elsewhere in the patient.

rt-PA is derived from a biological source, transformed CHO cells, and by definition is a biologic, not a drug. It is generally not possible to define biologics as discrete chemical entities or demonstrate a unique composition. Other biologics include blood fractionation products such as albumin and Factor VIII, and both live and killed viral vaccines.

The process used to make a biologic is closely monitored and regulated by regulatory agencies, because a significant change in the process may result in a product which is different from that previously reviewed and regulated, and hence may require a new license. Process changes made during the investigational new drug (IND) development stage, and before the license is approved, are more easily incorporated into a new product (from a regulatory point of view) than after the license is generated (15).

t-PA Production. Recombinant technology provides the only practical means of rt-PA production. The amount of t-PA required per dose is on the order of 100 mg. Cell lines of transformed CHO cells, selected for high levels of rt-PA expression using methotrexate, are grown in large fermenters (21). The purification steps for rt-PA must therefore separate out cells, virus, and DNA. The literature on the industrial practice of recovery and purification of rt-PA generated by suspension culture of Chinese hamster ovary cells is limited (15). Recovering a protein derived from mammalian cells involves a number of steps (15). One possible scheme is shown in Figure 2. The culture medium is separated from the cell by sterile filtration (see MICROBIAL AND VIRAL FILTRATION). This is followed by additional removal by cross-flow filtration, ultrafiltration (qv), and chromatography to remove DNA and remaining viruses. The product protein then undergoes purification by chromatography.

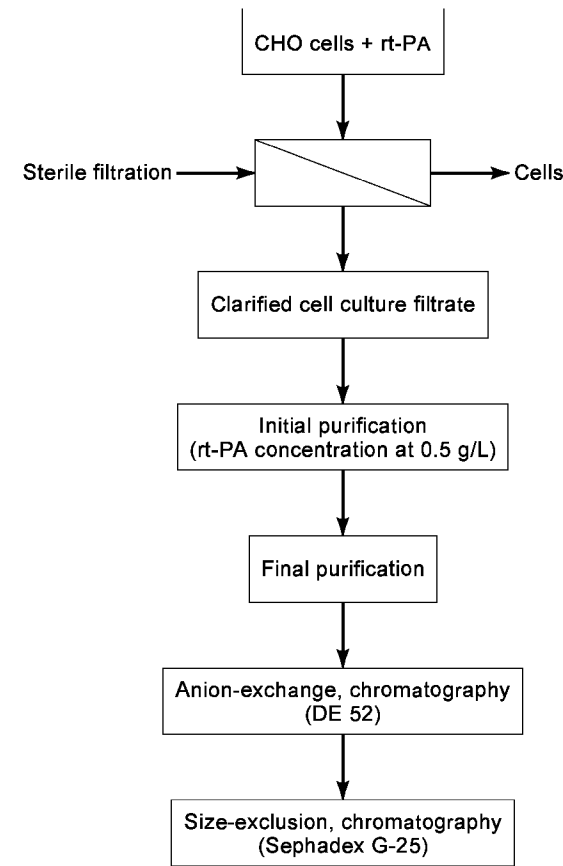


Fig. 2. Outline of possible steps in the recovery and purification sequence for recombinant tissue plasminogen activator derived from recombinant CHO cells.

The separation of cells from the culture media or fermentation broth is the first step in a bioproduct recovery sequence. Whereas centrifugation is common for recombinant bacterial cells (see CENTRIFUGAL SEPARATION), the final removal of CHO cells utilizes sterile-filtration techniques. Safety concerns with respect to contamination of the product with CHO cells were addressed by confirming the absence of cells in the product, and their relative noninfectivity with respect to immune competent rodents injected with a large number of CHO cells.

The possibility that DNA from recombinant immortal cell lines such as CHO cells could cause oncogenic (gene altering) events resulting in cancer (22) was a concern during development of the rt-PA purification sequence. Data suggest that DNA, by itself, is inactive *in vivo*; removal of the DNA, however, is still a concern. The goal for rt-PA purification is to reduce the DNA to <10 pg dose (10⁻¹¹ g dose). A level of 0.1 pg has been achieved, representing a ~9log reduction in the DNA, and requiring special assays to detect and quantify these very low levels of DNA in the final product (15).

Sensitive, specific, and if possible, rapid assays for product and potential contaminants are an essential part of separation methods selected for the purification sequence. Ultrafiltration (qv) followed by ion-exchange (qv) chromatography (qv) and then a final round of ultrafiltration concentrate the dilute protein while purifying it (23). Precipitation prior to chromatography could remove unwanted proteins before the sample is injected into the first liquid chromatography column. At the initial purification stage the rt-PA concentration is 0.5 g L; the DNA is at 0.11 ng mL. The use of anion-exchange chromatography (DE step) appears to be particularly effective in removing the DNA. Studies using another product, IgM, derived from cell culture showed DNA clearance may be enhanced by predigestion (hydrolysis) of the DNA using nucleases prior to the anionic-exchange chromatography step (24).

Independent Assays for Proving Virus Removal. Retroviruses and viruses can also be present in culture fluids of mammalian cell lines (15,24). Certainly the absence of virus can be difficult to prove. Model viruses, eg, NIH Rausher leukemia virus and NZB Xenotropic virus, were spiked into fluids being purified, and their removal subsequently validated when subjected to the same purification sequence as used for the product.

Viral clearance can be achieved by use of chaotropes such as urea or guanidine, pH extremes, detergents, heat, formaldehyde, proteases or DNA'ses organic solvents such as formaldehyde, or ion-exchange or size-exclusion chromatography. The protein product must be stable at the conditions used to deactivate or remove the virus. Because only the inactivation or removal which can be measured counts as validation, a sequence of orthogonal removal fractionation steps must be used (1,15,24). For a fluid spiked with 10⁷ virus particles mL, if the sensitivity of detection after each treatment is 10² particles mL, the analytical technique could only show a removal of 10⁴ particles mL. Hence to achieve evidence of 12 logs of clearance (10⁷–10⁻⁶), three different mechanisms of analysis would need to be used assuming each gives 4 logs of clearance (20). It is not valid to use the same approach, ie, the same step repeated three times, to achieve the 12 log reduction. Total clearance, based on the product of three separation steps (10⁴ × 10⁴ × 10⁴ = 10¹²), requires that the three steps be totally independent or orthogonal.

Another example of virus clearance is for IgM human antibodies derived from human B lymphocyte cell lines where the steps are precipitation, size exclusion using nucleases, and anion-exchange chromatography (24). A second sequence consists of cation-exchange, hydroxylapatite, and immunoaffinity chromatographies. Each three-step sequence utilizes steps based on different properties. The first sequence employs solubility, size, and anion selectivity; the second sequence is based on cation selectivity, adsorption, and selective recognition based on an anti-u chain IgG (24).

Purification of Human Cell-Line t-PA. The sequence of steps making up the initial and final purifications of recombinant t-PA from CHO cells is proprietary. A detailed experimental protocol for t-PA derived from normal, nonrecombinant human cultured cells (ATCC CRL-1459, American Type Culture Collection, Rockville, Maryland), however, is available (16). This provides insights into the types of chromatography steps which might be employed for purification of rt-PA. Human cell t-PA also contains urokinase plasminogen activator (u-PA) which, except for a single glycosylation (at Asn-302), is structurally similar to t-PA and tends to co-purify. The sequence in Figure 3 shows the steps for fractionating the two proteins. The yield is only 20 mg from 1400 L, illustrating the critical role of a recombinant cell line in obtaining both high yields and higher selectivity in producing a specific type of protein.

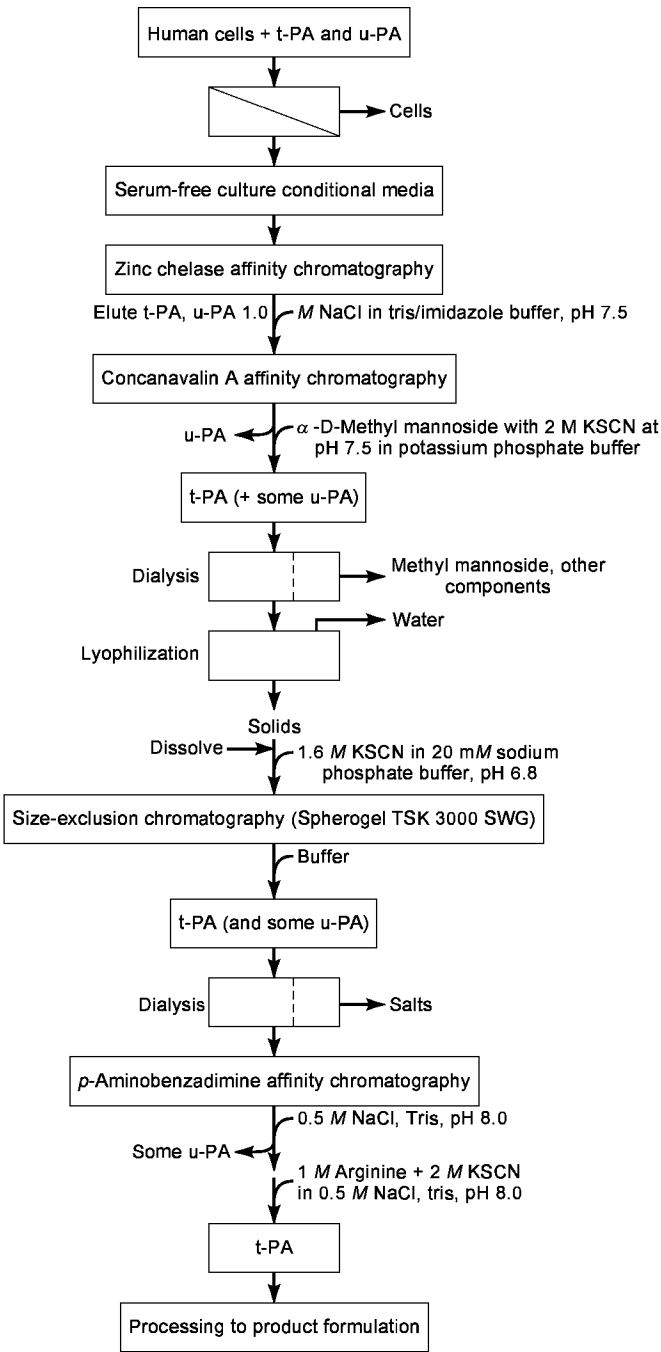


Fig. 3. Overview of purification sequence for the nonrecombinant tissue plasminogen activator (t-PA) which also contains urokinase plasminogen activator (u-PA). Serum-free culture conditional media is from normal human cell line. The temperature for all steps, except for size-exclusion chromatography (22°C), was 4°C. Adapted from Ref. 16.

Because the culture media contained both t-PA and u-PA, this separation required several extra affinity chromatography steps, as well as dialysis buffer exchange between the different chromatography columns (see Fig. 3). The salts and buffers added during the purification sequence must also be removed from the product at various points, adding significant complexity to the purification sequence. Desalting the buffer exchange constitute significant separation steps in the production of almost all biotechnology protein products.

Adsorption of t-PA to process equipment surfaces consisting of either stainless steel or glass was minimized by adding the detergent polyoxyethylene sorbitan monooleate (Tween 80) to the serum-free culture conditioned media at 0.01% (vol/vol). The equipment was also rinsed, before use, with phosphate buffered saline (PBS) containing 0.01% Tween 80. Hydrophilic, plastic equipment was used whenever possible. All buffers were sterile filtered. Sterile filtration of liquids and gases is usually carried out using 0.2 or 0.45 µm filters.

Manufacture of Biologics and Government Regulation

The difference between biologics and drugs is not only a matter of definition, it is also a process design issue. To compensate for the incomplete analytical capability to define biologics, regulatory agencies include parameters of the process used to make biologics in the control and monitoring. Changes in the process may yield a different product from that previously reviewed and approved. A different product requires a new license (15). Thus substantial barriers exist in terms of effort, money, and time to making significant changes in processes used to produce licensed biologics. Process changes are to be expected during the investigational new drugs (IND) phase and before the license is approved, but significant changes are rarely made after licensing. The time which can elapse between conception of an idea for a process change and granting of a new license can be as much as two years and cost several million dollars.

The definition of biologics versus drugs continues to evolve. Assignment is made on a case by case basis (25). Section 351 of the Public Health Service Act defines a biologic product as "any virus, therapeutic serum, toxin, antitoxin, vaccine, blood, blood component or derivative, allergenic product, or analogous product ... applicable to the prevention, treatment, or cure of diseases or injuries in man." Biologics are subject to licensing provisions that require that both the manufacturing facility and the product be approved. All licensed products are subject to specific requirements for lot release by the FDA. In comparison, drugs are approved under section 505 of the FD & C act (21 USC 301–392), where there is not lot release by the FDA except for insulin products. Insulin, growth hormone, and many other hormones have been treated as drugs, whereas erythropoietin (EPO), which also fulfills the criteria of a hormone, was reviewed in the biologic division of the FDA. Insulin is derived from a bacterial fermentation; EPO is obtained from mammalian cell culture. Hormones, for the most part, are expected to be reviewed as drugs.

The design of bioseparation unit operations is influenced by these governmental regulations. The constraints on process development grow as a recovery and purification scheme undergo licensing for commercial manufacture.

Protein Chromatography

Proteins and nucleotides are macromolecular biomolecules. Mixtures of biomolecules are fractionated based on differences in charge; molecular weight, shape, and size; solubility in organic solvents; surface hydrophobic character; and types of active sites using ion-exchange, size-exclusion (gel-permeation), and reversed-phase, hydrophobic interaction (surface hydrophobicity), and affinity chromatographies, respectively. The appropriate separation may be selected from these five basic classes of chromatography. More than 30% of the purification steps for laboratory-scale protein purification procedures use ion-exchange and or gel filtration, and at least 20% use affinity chromatography (23). This pattern is likely to be consistent with industrial practice. The following represent some chromatography column options for biopharmaceutical proteins (10). Sepharose, Sephadex, and Sephacryl resins are supplied by Pharmacia; Spherox, Spherosil, and Trisacryl resins are supplied by Sepracor, Inc.; Toyopearl resins are supplied by TosoHaas; Fractogel resins are supplied by E. Merck Separations; Bakerbond resins and silicas are supplied by J. T. Baker. For adsorption, silica may be used.

Ion exchange

DEAE Sepharose Fast Flow LC
DEAE Sepharose Fast Flow HC
DEAE Spherox M
DEAE Spherosil M
DEAE Trisacryl Plus M
Toyopearl DEAE-650 (M)
Fractogel EMD DMAE-650 (M)
Fractogel EMD DEAE-650 (M)
Q Sepharose Fast Flow
QMA Spherox M
QMA Spherosil M
QMA Trisacryl Plus M
Toyopearl QAE-550 C
SP Sepharose Fast Flow
SP Sepharose High Performance
SP Sepharose Big Bead
SP Trisacryl Plus M
Toyopearl SP-650 M
Fractogel EMD SO₃-650 M

Gel permeation

Agarose, 16%
Sephadex G25 Medium
Sephadex G75
Trisacryl Plus GF 03 M
Trisacryl Plus GF 10 M
Trisacryl Plus GF 20 M
Sephacryl S-100HR
Sephacryl S-200HR
Toyopearl HW-50F
Toyopearl HW-55F

Hydrophobicity

Toyopearl Bulyl-650 M
Bulyl Spherox M
Toyopearl Ether-650 M
Octyl Spherox M
Phenyl Spherox M
Toyopearl Phenyl-650 M
Bakerbond Hi-Propyl

Affinity

Blue Sepharose CL-6B
Red Sepharose CL-6B
Blue Spherox M
Baseline Blue Trisacryl M
Heparin Sepharose CL-6B
Heparin Spherox M
Toyopearl AF-Chelate-650 M (Copper)
Toyopearl AF-Chelate-650 M (Zinc)

Reversed-phase chromatography is widely used as an analytical tool for protein chromatography, but it is not as commonly found on a process scale for protein purification because the solvents which make up the mobile phase, ie, acetonitrile, isopropanol, methanol, and ethanol, reversibly or irreversibly denature proteins. Hydrophobic interaction chromatography appears to be the least common process chromatography tool, possibly owing to the relatively high costs of the salts used to make up the mobile phases.

Liquid Chromatographs. The basic equipment for liquid chromatography is shown in Figure 4. The column is packed with an adsorbent, ie, the stationary phase. The mixture to be separated is pushed through the column by the eluent or mobile phase (26). Isocratic chromatography, carried out at a constant flow rate, buffer composition, and temperature, is usually associated with size-exclusion separations. Gradient chromatography typically uses a constant flow rate and temperature, but the composition of the element is altered by mixing two or more buffer reservoirs to achieve a steadily changing salt concentration or changes in pH. The gradients formed are reported in terms of concentration at the inlet of the chromatography column; protein is detected at the column outlet. A chromatogram of the type illustrated in Figure 5 results.

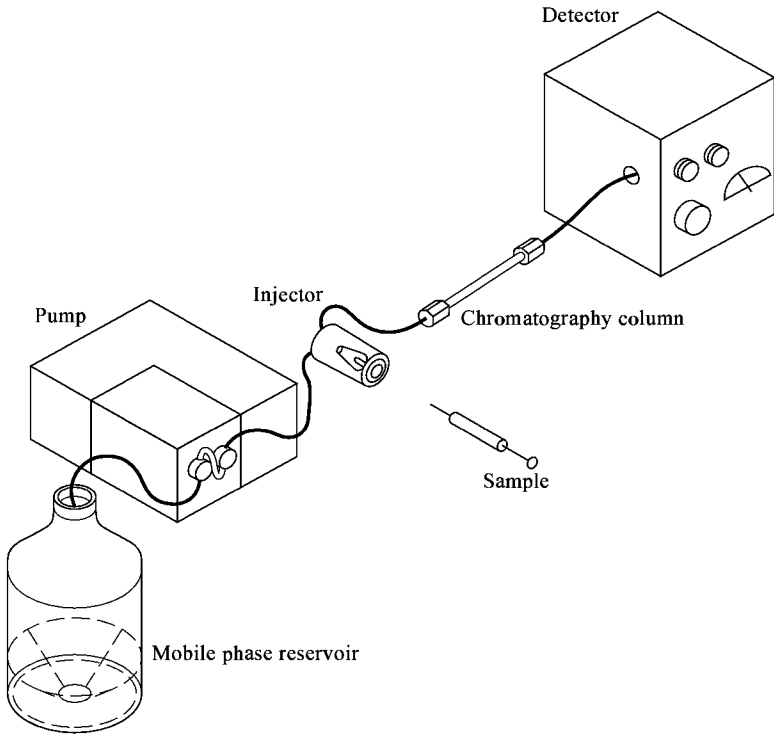


Fig. 4. Schematic of a process liquid chromatography system (16).

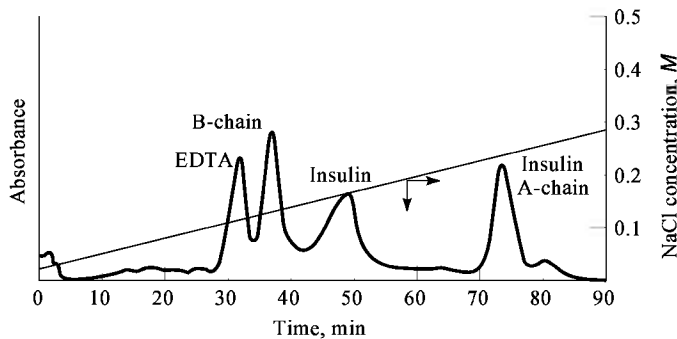


Fig. 5. Anion-exchange separation of insulin, and insulin A- and B-chains, over diethylaminoethyl (DEAE) in a 10.9 × 200 mm column having a volume of 18.7 mL. Sample volume is 0.5 mL and protein concentration in 16.7 mM Tris buffer at pH 7.3 is 1 mg/mL for each component in the presence of EDTA. Eluent (also 16.7 mM Tris buffer, pH 7.3) flow rate is 1.27 mL/min, and protein detection is by uv absorbance at 280 nm. The straight line depicts the salt gradient.

Courtesy of the American Chemical Society (48)

The concentration profile of the gradient at the outlet of the column can be significantly different from the profile at the inlet when a component (referred to as a modulator) in the eluting buffer adsorbs onto the stationary phase in a nonideal manner causing the gradient to deform as it passes through the column. What appears to be anomalous peak behavior can result including self-concentration of a peak and the appearance of shoulders or multiple peaks for single sample components known to be homogeneous. This can occur for gradients used in reversed-phase, ion-exchange, or affinity chromatography (27–29).

Ion-Exchange Chromatography. Ion-exchange chromatography is initiated by eluting an injected sample through a column using a buffer but no NaCl or other displacing salt. The protein, which has charged sites spread over its surface, displaces anions or cations previously equilibrated on the stationary phase, ie, the protein sites exchange with the salt counterions associated with the ion-exchange stationary phase. A protein having a greater number and/or density of charged sites displaces or exchanges more ions and hence binds more strongly than a protein having a lower charge number or charge density.

Proteins deform and change shape in response to the environment. Hence, a protein left on the surface of an ion-exchange resin for a day or longer may slowly start to unfold exposing an increasing number of charged sites to bind with the ion-exchange resin. It is possible that this process can continue until the protein binds so strongly that it is impossible to desorb the protein without dissolving it, in NaOH, for example, and destroying it. To prevent such a situation, ion-exchange chromatography must be completed in a matter of hours or less.

After the column is loaded, proteins of similar size and shape are separated by differential desorption from the ion exchanger by using an increasing salt gradient of the mobile phase. The more weakly bound macromolecules elute first; the most tightly bound elute last, at the highest salt concentration. Figure 5 is an example of an anion-exchange separation (48). Prior to injection of the sample, the column was equilibrated with the 16.7 mM Tris buffer; the EDTA stabilized the solubility of the insulin sample injected onto the column. All of the proteins are initially retained on the anion-exchange stationary phase (DEAE 650 M) during loading of the sample onto the column. Subsequent application of the NaCl gradient, formed by the controlled mixing of buffers from two reservoirs of mobile phase, elutes the proteins. One reservoir contains only the 16.7 mM Tris buffer; the second contains 0.5 M NaCl in the same buffer. Following the elution of the last peak, the column may be flushed using a buffer at a high (2.5 M NaCl) salt concentration to verify that all proteins are desorbed. In some cases, a cleaning procedure is performed by passing methanolic NaOH through the column. The column is then re-equilibrated using the salt-free buffer, by pumping approximately 10 column volumes of the buffer through the stationary phase, or until the pH of the effluent and influent are the same to prepare the column for another injection.

Amphoteric Properties Determine Conditions for Ion-Exchange Chromatography. Proteins, amphoteric polymers of acidic, basic, and neutral amino acid residues, carry both negatively and positively charged groups on the surface, the ratio depending on pH (30). The isoelectric point (pI) is the pH at which a protein has an equal number of positive and negative charges. Proteins in solutions at a pH > pI have a net negative charge. Below the pI, proteins have a net positive charge. Many proteins have a pI < 7 and are processed using buffers having a pH of 7 to 8. Thus anion exchangers (positively charged stationary phases) are popular for protein chromatography. Ion-exchange matrices derivatized having negatively charged groups are cation exchangers. These bind positively charged proteins, ie, cations, when the mobile phase pH is < pI.

The selection of the pH of the buffer, as well as the type of ion-exchange (anion or cation) stationary phase is a function of the amphoteric nature of the protein and protein stability as a function of pH. For example, for a protein stable at pH > 6.5, having pI = 5.5, an anion exchanger is appropriate when the separation is run in a buffer of pH > 6.5. If this protein were stable at pH = 5, and the pI = 5.5, a cation exchanger and buffer of pH < 5.0 would be appropriate.

The ion-exchange (qv) groups used to derivatize stationary phases for the purification of proteins are summarized in Table 2. Corresponding buffers are given in Table 3. Strong anion and cation exchangers are almost fully ionized at pH = 3–11 and coincide with the pH range of protein purification. Weak anion and cation exchangers have a narrower pH range over which they are ionized. Anion exchangers are preferred because desorption of the protein is more readily accomplished at lower salt concentrations.

Table 2. Ion-Exchange Groups Used in Protein Purification^a

Name	Abbreviation	Formula
<i>Weak anion</i>		
aminoethyl	AE	—C ₂ H ₄ NH ₃ ⁺
diethylaminoethyl	DEAE	C ₂ H ₄ NH(C ₂ H ₅) ₂ ⁺
<i>Weak cation</i>		
carboxy	C	—COO [−]
carboxymethyl	CM	CH ₂ COO [−]
<i>Strong anion</i>		
trimethylaminoethyl	TAM	CH ₂ N(CH ₃) ₃ ⁺
triethylaminoethyl	TEAE	C ₂ H ₄ N(C ₂ H ₅) ₃ ⁺
diethyl-2-hydroxypropylaminoethyl	QAE	C ₂ H ₄ N ⁺ (C ₂ H ₅) ₂ CH ₂ CH(OH)CH ₃
<i>Strong cation</i>		
sulfo	S	SO ₃ [−]
sulfomethyl	SM	CH ₂ SO ₃ [−]
sulfopropyl	SP	—C ₃ H ₆ SO ₃ [−]

^a Courtesy of IRL Press (30).

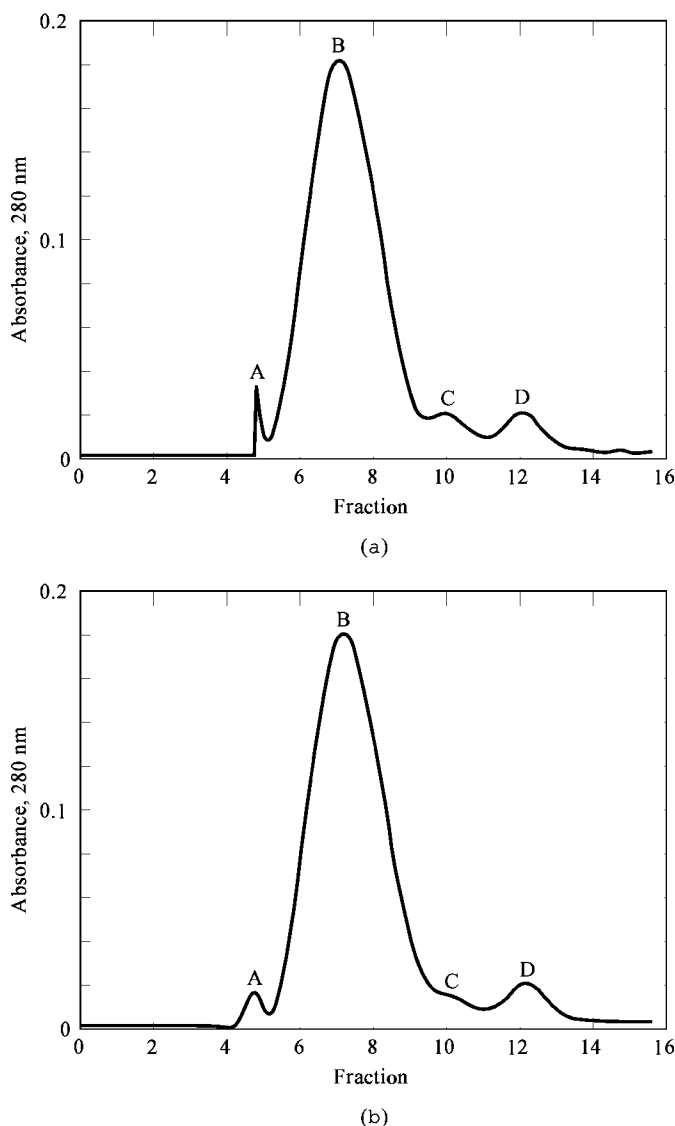


Fig. 7. Chromatograms of size-exclusion separation of IgM (mol wt = 800,000) from albumin (69,000) where A–D correspond to IgM aggregates, IgM, monomer units, and albumin, respectively, using (a) FPLC Superose 6 in a 1 × 30 cm long column, and (b) Sepharose CL-6B in a 37-cm column.

Courtesy of American Chemical Society (24).

Buffer Exchange and Desalting. A primary use of size-exclusion chromatography (sec) is for removal of salt or buffer from the protein, ie, desalting and buffer exchange (32). The difference in molecular weights is large; salts generally have a mol wt < 200, whereas mol wts of proteins are between 10,000 and 60,000.

Alternative methods of desalting and buffer exchange include continuous diafiltration, countercurrent dialysis (ccd), a membrane separation technique, and cross-flow filtration, which uses membranes (see MICROBIAL AND VIRAL FILTRATION). Both of those methods rely on filtration at a molecular scale, using membranes having porosities which reject proteins but allow passage of salts. Membrane methods are often preferred for an unspecified protein because these procedures are less costly and enable higher throughput than size-exclusion chromatography. Buffer exchange, used to remove denaturing agents in order to induce refolding of proteins, to remove buffers between purification steps, or to remove buffers and other reagents from the final product, is usually carried out at later steps in a recovery sequence (see Figs. 2 and 3). Equations for calculating separation efficiencies and recovery yields for all three methods for a specific case study using a recombinant protein of 160,000 mol wt are available (32). Size-exclusion chromatography using Sephadex G-25M gel-filtration media in this case had disadvantages compared to the membrane filtration techniques giving 100 × less complete ion removal, 130–1200% greater dilution, 30% higher cost, 66% higher (eluting) buffer requirement, 50% higher space on the plant floor, 50% higher operating time, and half the throughput. However, cross- or tangential-flow filtration (tff) required up to 90 passes through a pump whereas sec and ccd were single pass. Other disadvantages of tff include frequent changeout of membranes and relatively large volumes of protein feed being required for laboratory-scale tests, a particular disadvantage for recombinant products. The more recent testing of a novel size-exclusion stationary phase, however, which facilitates rapid preparative (process-scale) separation of salts from protein in less than seven minutes, shows that process size-exclusion chromatography is capable of high throughput while reducing the volume needed to obtain proper refolding of the protein. Salts causing the protein to be denatured were rapidly removed and reduced to a level where protein refolding occurred (33). The use of sec is likely to continue to be a widely practiced technology in industry. Rapid size-exclusion columns for the purpose of buffer exchange have been developed which enable desalting to be achieved at linear velocities of 500–600 cm/h (33), significantly increasing throughput and reducing operating time and plant floor space. Further, sec using gel-filtration media on cellulosic-based material has a special niche for the partial and controlled separation of denaturing salts from recombinant proteins for purposes of refolding. The development of such rapid desalting techniques is important because of the larger volumes of proteins needing to be processed in industry.

Column Size. Size-exclusion chromatography columns are generally the largest column on a process scale. Separation is based strictly on diffusion rates of the molecules inside the gel particles. No proteins or other solutes are adsorbed or otherwise retained owing to adsorption, thus, significant dilution of the sample of volume can occur, particularly for small sample volumes. The volumetric capacity of this type of chromatography is determined by the concentration of the proteins for a given volume of the feed placed on the column.

The volume of the solvent between the point of injection and the peak maximum of the eluting protein is defined as the elution volume, V_e (Fig. 8). The fluid volume between the particles of the stationary phase is the extraparticulate void volume or exclusion volume, V_o . The porosity of the stationary phase, determined by the extent of cross-linking of the polymers which make up the particles of a gel-permeation matrix, determines the extent to which a protein or other solute can explore the intraparticulate void volume. The higher the cross-linking, the smaller the effective pore size and the lower the molecular weight (or size) of the molecule which is excluded from the gel. Hence, the apparent porosities of a gel-permeation column are a function of the molecular probes used to measure it. For a large molecule that is completely excluded, the void volume is equivalent to the extraparticulate void volume, V_o . Molecules that are small enough to penetrate the gel have an elution volume $> V_o$. A small molecule, such as a salt, can potentially explore almost all of the

bed volume and has the following elution volume:

$$V_e = V_o + K_d V_s$$

(2)

where K_d represents the fraction of the volume of the mobile phase inside the particle, V_s , which can be explored by the molecular probe. For a probe small enough to explore all of the intraparticle void volume, K_d is 1 and the elution volume is $V_o + V_s$. Because the combined volume of the fluid between the particles, V_o , and inside the particles, V_s , cannot exceed the total volumes of the column, $V_o + V_s$ must be less than column volume V_t . All components injected into a size-exclusion column thus elute in one column volume.

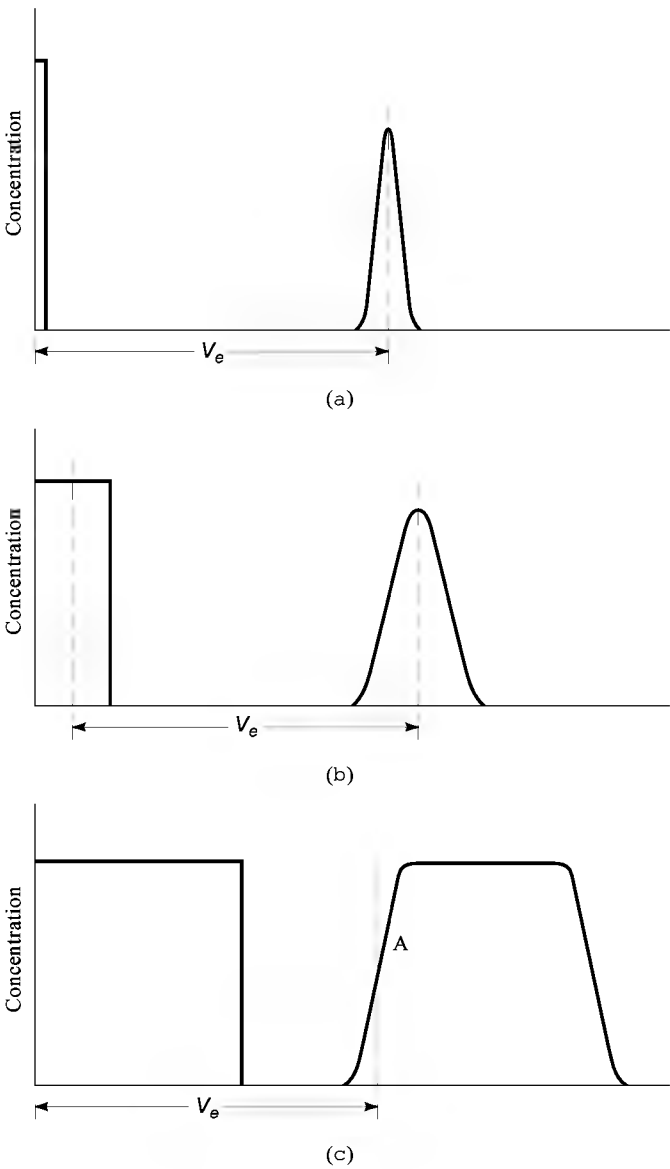


Fig. 8. Representation of measurement of elution volume, V_e , as a function of sample volume: (a) <2% of bed volume, (b) >2% and (c) >2% and giving a plateau region which has the same concentration as the injected sample; A represents the inflection point. See text.

Courtesy of Pharmacia (34).

The total stationary-phase volume required to process a given feed stream is proportional to the inlet concentration and volume of the feed. For example, for a typical inlet concentration of protein of 10 g/L, in a 100 L volume of feed, a column volume of at least 100 L is needed for size-exclusion chromatography. In comparison, an ion-exchange column having an adsorption capacity of 50 g/L would only require 20 L of column volume for the same feed.

Elution and Sample Volumes. The elution volume, V_e , is measured from the beginning of the injection to the center of the peak maximum for a Gaussian peak, if the sample volume is negligible relative to the elution volume (see Fig. 8a). A negligible injection volume is defined as being <2% of the elution volume. The elution volume for samples larger than this are measured from the halfway point of the volume injected to the center of the eluting peak (see Fig. 8b). In samples which are so large that a plateau region is obtained having the same concentration as the sample (see Fig. 8c), the elution volume is measured from the start of the sample injection to the inflection point of the peak.

Sample dilution in gel permeation can be 10-fold or more for small volumes. Hence, proper representation of the inlet concentration (eg, see Fig. 8a), would require that the injection pulse be much higher because the area under the eluting peak and under the injected sample should be the same. Similarly, dilution of the injected sample would occur in the cases represented by Figures 8b and 8c, although the difference in heights between injected pulse and the maximum of the eluting peak would be smaller, because diffusion of the solute away from the center of mass occurs at the leading (left side) and trailing (right side) edges of the peak. If the peak is broad enough, ie, the sample volume is large enough, the solute at the center of the peak is not diluted owing to diffusion and the peak maximum is therefore equal to that of the injected sample.

Distribution Coefficients. Gel-permeation stationary-phase chromatography normally exhibits symmetrical (Gaussian) peaks because the partitioning of the solute between mobile and stationary phases is linear. Criteria more sophisticated than those represented in Figure 8 are seldom used (34).

The elution volume, V_e , and therefore the partition coefficient, K_D , is a function of the size of solute molecule, ie, hydrodynamic radius, and the porosity characteristics of the size-exclusion media. A protein of higher molecular weight is not necessarily larger than one of lower molecular weight. The hydrodynamic radii can be similar, as shown in Table 4 for ovalbumin and α -lactalbumin. The molecular weights of these proteins differ by 317%; their radii differ by only 121% (53).

Table 4. Properties of Standard Proteins^a

Protein	Mol wt	pI	Radius, nm	Asymmetry, f/fo ^b
bovine serum albumin	66,000	5.74	3.50	1.29
ovalbumin	45,000	5.08	2.78	1.16
α-lactalbumin	14,200	4.57	2.30	1.18
myoglobin	16,900	7.10	2.40	1.18

^a Adapted from Ref. 53.
^b Frictional coefficients of f, solvated protein, and fo, nonsolvated sphere.

Some types of size-exclusion phases, based on silica or macroporous polymeric materials, having rigid pores and defined pore size distributions. The dominant types of gels used in industry are dextran cross-linked epichlorophydran (Sephadex), agarose prepared from agar (Sephacrose), or allyl dextran cross-linked with *N,N'*-methylenebisacrylamide (Sephacryl). These materials imbibe significant quantities of water and have bed volumes ranging from 2 to 3 mL/g dry weight of stationary phase for Sephadex G-10 (nominal molecular weight cutoff of 10,000) to 20–25 mL/g dry weight of stationary-phase Sephadex G-200 (nominal molecular weight cutoff of 200,000). Their structures resemble a cross-linked spider web, where the extent of cross-linking or association between hydrated polymer chains, rather than specific pore sizes, determine the apparent pore size distribution. The hydrophilic character of the polymers which make up these gels require cross-linking to prevent dissolution. The hydrophilic character is compatible with the majority of industrially relevant proteins, most of which can be denatured by hydrophobic surfaces but preserved in active confirmation at hydrophilic conditions. This property can be offset by poor flow properties, however, particularly for lightly cross-linked gels, because these gels are soft and have a tendency to compress when flow rates exceed a threshold which decreases with decreasing extents of cross-linking. Hence, Sephadex G-10 has the highest cross-linking and flow stability, and the lowest specific bed volume, but also has the lowest effective pore size or porosity, limiting its sieving capabilities. Sephadex G-200 has the lowest cross-linking and flow stability and the highest specific bed volume and effective pore size. Sephadex 200 is useful for separating high molecular weight proteins, but at relatively low flow rates (Table 5).

Table 5. Comparison of Gel-Permeation Stationary Phase^a

Sephadex G-X ^b	Specific volume water mL/g dry gel	Permeability, <i>K</i> _o	Operating pressure ^c , kPa ^d	Flow rate ^c water, mL/(cm ² ·h)
10	2–3	19	f	f
15	2.5–3.5	18	f	f
25	4–6	9–290 ^e	f	f
50	9–11	13.5–400 ^e	f	f
75	12–15		160	77
100	15–20		96	50
150	20–30		36	23
200	30–40		16	12

^a Adapted from Ref. 34.
^b Corresponds to the nominal cutoff value for wt × 10³, eg, G-10 has a mol wt cutoff value of ~10,000.
^c Value is maximum unless otherwise noted.
^d To convert kPa to cm H₂O, multiply by 10.2.
^f May be calculated using Darcy's law: *U* = *K*_o (Δ*P*/*L*), where *U* is linear flow as mL/(cm²·h), *L* is bed length in cm, Δ*P* is pressure drop over gel bed in kPa^d, and the maximum pressure is 30.4 kPa^d (310 cm H₂O).
^e Depends on particle size (dp); as dp increases, *K*_o increases.

The distribution coefficient, *K*_o, represents the fractional volume of a specific stationary phase explored by a given solute, represented by equation 3:

$$K_d = \frac{V_e - V_o}{V_s}$$

(3)

where *V*_o is the void volume, *V*_i is the volume of the solvent (usually aqueous buffer) inside the gel which is available to very small molecules, and *V*_e is the elution volume of a small volume of injected molecular probe. The measurement of *V*_i is difficult, requiring use of an ion or small molecule which freely diffuses into all of the fluid volume inside the gel particles and then is readily detected at the outlet of the column. D₂O and radioactive ²³Na have been used. The latter is detected by a refractive index detector. An indirect measurement of *V*_i is more convenient and adequate. The column void volume (Fig. 9) may be measured using a soluble, high molecular weight target molecule which, because it does not explore any of the internal fluid volume of the stationary phase, it only distributed in the mobile phase. Blue dextran, a water-soluble, sulfonated, blue-colored dextran having mol wt > 669,000, manufactured by Pharmacia, and DNA (Type III from salmon tests) have been employed (26). The total column volume, *V*_s, can be calculated from the dimensions of the bed, although the direct measurement of column volume using water displacement before packing is more accurate. (It should be noted that total column volume is also represented by *V*_i in the recent literature). The difference, *V*_i – *V*_o, is then taken as an approximation of *V*_i. On this basis, *K*_{av}, the fraction of stationary phase volume available for a given solute species, is defined as in equation 4:

$$K_{av} = \frac{V_e - V_o}{V_i - V_o}$$

(4)

The constant *K*_{av}, is not a true partition coefficient because of difference, *V*_i – *V*_o, includes the solids and the fluid associated with the gel or stationary phase. By definition, *V*_i represents only the fluid inside the stationary-phase particles and does not include the volume occupied by the solids which make up the gel. Thus *K*_{av} is a property of the gel, and like *K*_d it defines solute behavior independently of the bed dimensions. The ratio of *K*_{av} to *K*_d should be a constant for a given gel packed in a specific column (34).

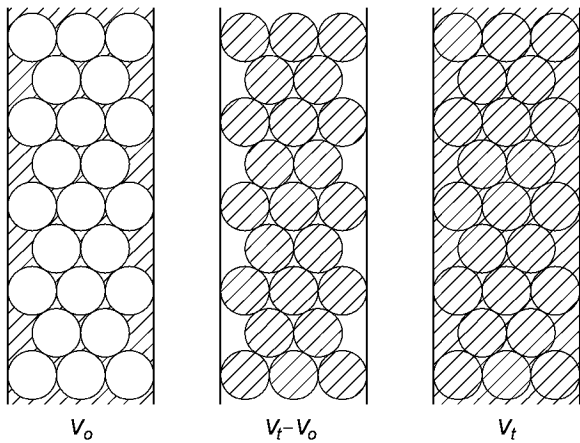


Fig. 9. Diagrammatic representation of V_t and V_o . $V_t - V_o$ includes the volume of the solid material forming the matrix of each bead.

Courtesy of Pharmacia (34).

Selectivity curves result from measured values of K_{av} plotted vs log (mol wt) enabling molecular weight determination of globular proteins having similar asymmetry factors. A sphere has an asymmetry factor of 1; an ellipsoid has a factor > 1 (51). Such curves are linear (Fig. 10), the intercept increasing with increasing porosity (decreasing cross-linking). Extrapolation of these curves through the x -axis yields the molecular weights of probes which should be completely excluded from the gels because these target molecules are larger than the largest pores. Theoretically, the K_{av} for a given molecular probe should have a value between 0 and 1. A completely excluded molecule has $K_{av} = 0$; molecules able to completely explore the fluid inside the stationary-phase particle have $K_{av} = 1$. If K_{av} is less than zero, then channeling owing to a poorly packed bed is a probably cause. If the K_{av} is greater than 1, an interaction (adsorption) of the molecular probe with the stationary phase is a likely explanation.

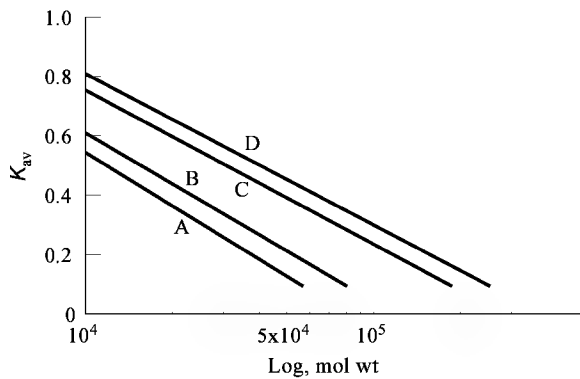


Fig. 10. Selectivity curves A–D for Sephadex G-75, G-100, G-150, and G-200, respectively, for globular proteins.

Courtesy of Pharmacia (34).

Gel-permeation media are extremely versatile and may be used for separation of particles such as viruses (Fig. 11) as well as proteins (34). Separations of proteins and other particles having sizes equivalent to a molecular weight of 40×10^6 are possible using the agar-based Sepharose-type gel. This particular gel has a limited temperature range for operation, however. It melts upon heating to 40°C (34).

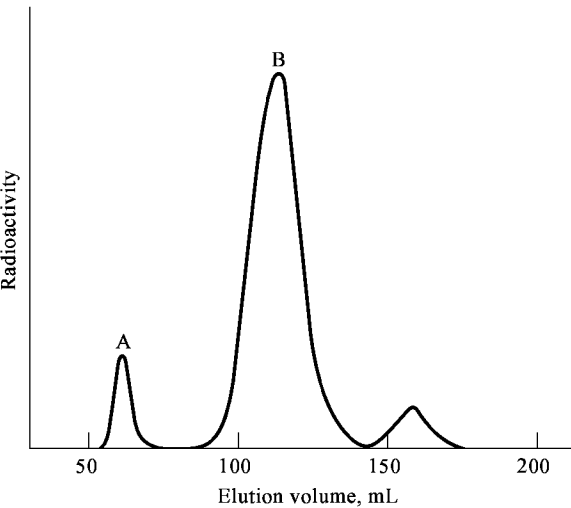


Fig. 11. Separation of ^{32}P -labeled A, adenovirus Type 5, and B, poliovirus Type 1 on Sepharose 2B where the column is 2.1×56 cm; eluent, 0.002 M sodium phosphate, pH 7.2, containing 0.15 M NaCl; and flow rate, 2 mL (cm²h).

Courtesy of Pharmacia (34).

The gels having a larger extent of cross-linking, particularly Sephadex G-10, G-15, and G-25, may retain through weak adsorption some types of aromatic molecules, and consequently impart reversed-phase properties to the gel. This may be the result of the weakly hydrophobic character of the cross-linking agents used in the synthesis of these gels. A recently introduced product, based on crosslinked agarose combined with crosslinked dextran (Superdex), exhibits low nonspecific interactions as well as enhanced resolution (34). An excellent overview of gel permeation chromatography is given in

Reference 51.

Reversed-Phase Chromatography. Stationary phases for reversed-phase chromatography consist principally of silica particles, silica supports having a hydrocarbon bonded phase, or polymeric materials based on either vinyl or styrene–divinylbenzene copolymers (35–41). Mobile phases commonly used in reversed-phase chromatography are aqueous methanol, 2-propanol, and acetonitrile. These solvents are often mixed with acidic buffers containing small amounts of acids such as trifluoroacetic acid or hexafluorobutyric acid. These acids reduce the pH of the mobile phase to 3 and give sharper peaks by suppressing ionization of silanol groups of silica-based reversed-phase supports, and minimizing ionic effects (42).

The most prevalent type of stationary phase is made of silica or another type of inorganic support derivatized and bonded with an octadecyl (C₁₈) or octyl (C₈) coating. Nonderivatized silicas are sometimes utilized for process chromatography. Much like ion-exchange chromatography, the organic component, sometimes referred to as a modifier, is mixed with water or buffer to form an increasing gradient of the modifier. This gradient serves to elute the components, which are initially adsorbed onto the stationary phase in order of increasing hydrophobic character. Methanol is used as the modifier for eluting weakly adsorbed, hydrophilic peptides; 2-propanol is used to elute strongly adsorbed, hydrophobic peptides. Acetonitrile is widely used for separations of proteins and many other types of molecules because this solvent exhibits favorable mass-transfer properties, lower viscosity (and back-pressure) than the other solvents, and good eluting strength. Methanol has a higher aqueous heat of mixing than the other solvents, and this can lead to solvent degassing and bubble formation in the column. Bubbles interfere both with the operation of the column, causing peak dispersion, and detection of the peaks at the column outlet. Bubbles give anomalous peaks in spectrophotometric detectors and refractive index detection (42).

Product Monitoring and Peptide Mapping. Reversed-phase chromatography is widely used for analysis of proteins. Historically, the principal use of reversed-phase chromatography has been in the analysis and process separations of peptides, amino acids, and organic compounds which are characterized by lower molecular weight and solubility in acetonitrile or alcohol gradients. These mobile phases denature proteins and some polypeptides. Hence, reversed-phase chromatography is infrequently used for process-scale purification of proteins. Rather the excellent protein resolving power of this type of chromatography is employed on an analytical scale using columns packed with 2–10 mL of stationary phases having 1–5 µm particle sizes and for sample volumes which typically range from 1 to 10 µL.

One purpose in monitoring a protein product is to detect the presence of a change in which as little as one amino acid has been chemically or biologically altered or replaced during the manufacturing process. Variant amino acid(s) in a protein may not affect protein retention during reversed-phase chromatography if the three-dimensional structure of the polypeptide shields the variant residue from the surface of the reversed-phase support (20). Reversed-phase chromatography discriminates between different molecules on the basis of hydrophobicity. Because large proteins may contain only small patches of hydrophobic residues, these patches may not correlate to the molecular modifications which a reversed-phase analytical method seeks to detect. The reversed-phase method must therefore be completely validated, and preferably combined with controlled chemical and or proteolytic hydrolysis followed by chromatography or electrophoresis (see ELECTROSEPARATIONS) of the cleared protein to give a map of the resulting peptide fragments (20,43).

A peptide map is generated by cleaving a previously purified protein using chemicals or enzymes. Hydrolytic agents having known specificity are used to perform limited proteolysis followed by resolution and identification of all the peptide fragments formed. Identification of changes, and reconstruction of the protein's primary structure, is then possible. Reagents and enzymes which cleave specific bonds are discussed in the literature (44).

An example of a peptide map generated by trypsin hydrolysis of recombinant tissue plasminogen activator (rt-PA) is shown in Figure 12. The chromatogram shows the resolving power of reversed-phase high performance liquid chromatography in separating peptides obtained from t-PA in which the disulfide bonds had been reduced and alkylated prior to enzyme hydrolysis. The small peptides formed have little or no three-dimensional structure. Hence, measurable shifts in elution profiles occur when there are variant amino acids because a single amino acid change in a peptide has a larger effect on its solubility and retention than the same change has in a protein. The replacement of arginine at position 275 in a normal t-PA molecule with glutamic acid results in a significant peak shift (see Fig. 12) (43), showing how tryptic mapping can be a suitable method for monitoring lot-to-lot consistency of this particular recombinant product (20).

Reversed-phase high performance liquid chromatography has come into use for estimating the purity of proteins and peptides as well. However, before employed, a high performance liquid chromatographic (hplc) profile of a given protein must be completely validated (43).

Insulin Purification. An example of the purification of recombinant product by reversed-phase chromatography is recombinant insulin, a polypeptide hormone. Insulin consists of 51 amino acid residues in two chains and is relatively small. Reversed-phase chromatography is used after most of the other impurities have been removed by a prior ion-exchange step (see Fig. 1) (12). The method utilizes a process-grade C₈ reversed-phase support (Zorbex) having a particle size of 10 µm (14). Partially purified insulin crystals, dissolved in a water-rich mobile phase, are applied to the column and then eluted in a linear gradient generated by mixing 0.25 M aqueous acetic acid to 60% acetonitrile. The acidic mobile phase gives excellent resolution of insulin from structurally similar insulin-like components. The ideal pH is from 3.0 to 4.0, below insulin's isoelectric point, pI = 5.4. Under mildly acidic conditions insulin may deamidate to monodesamido insulin, but if the reversed-phase separation is done within a matter of hours, the deamidation can be minimized.

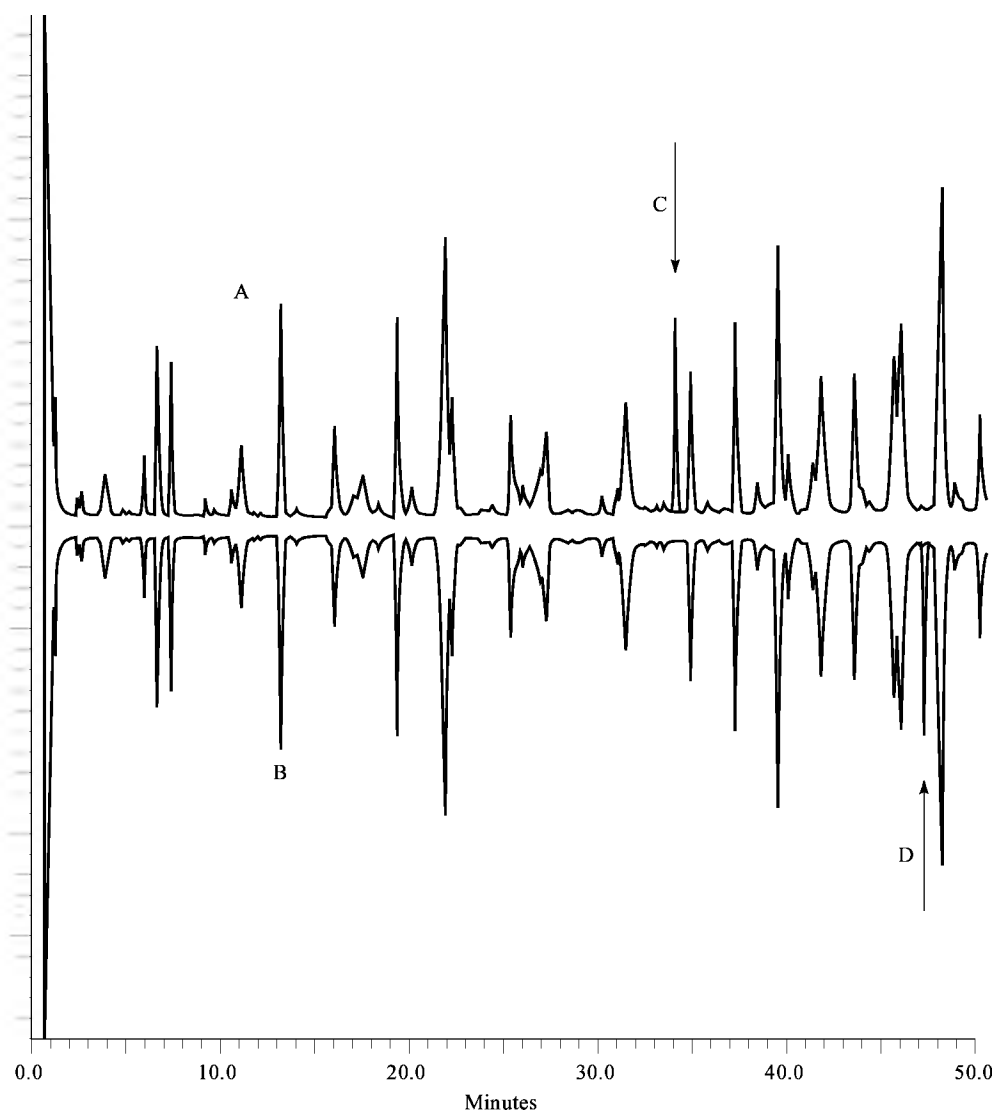


Fig. 12. Tryptic map of rt-PA (mol wt 66,000) showing peptides formed from hydrolysis of reduced, alkylated rt-PA. Separation by reversed-phase octadecyl (C_{18}) column using aqueous acetonitrile with an added acidic agent to the mobile phase. Arrows show the difference between A, normal, and B, mutant rt-PA where the glutamic acid residue, D, has replaced the normal arginine residue, C, at position 275.

Courtesy of Marvel Dekker (43).

This reversed-phase chromatography method was successfully used in a production-scale system to purify recombinant insulin. The insulin purified by reversed-phase chromatography has a biological potency equal to that obtained from a conventional system employing ion-exchange and size-exclusion chromatographies (14). The reversed-phase separation was, however, followed by a size-exclusion step to remove the acetonitrile eluent from the final product (12,14).

Whereas recombinant proteins produced as inclusion bodies in bacterial fermentations may be amenable to reversed-phase chromatography (42), the use of reversed-phase process chromatography does not appear to be widespread for higher molecular weight proteins.

Reversed-Phase Process Chromatography. Polypeptides, peptides, antibiotics, alkaloids, and other low molecular weight compounds are amenable to process chromatography by reversed-phase methods. There are numerous examples of bioproducts which have been purified using reversed-phase chromatography. The manufacture of salmon calcitonin, a 32-residue peptide used for treatment of post-menopausal osteoporosis, hypercalcemia, and Paget's disease of the bone, includes reversed-phase chromatography. This peptide, commercially prepared on a kilogram scale by a solid-phase synthesis, is then purified by a multimodal purification train. Reversed-phase chromatography is the dominant technique used by Rhône-Poulenc Rorer (45).

Another example is the purification of a β -lactam antibiotic, where process-scale reversed-phase separations began to be used around 1983 when suitable, high pressure process-scale equipment became available. A reversed-phase microparticulate (55–105 μm particle size) C_{18} silica column, with a mobile phase of aqueous methanol having 0.1 M ammonium phosphate at pH 5.3, was able to fractionate out impurities not readily removed by liquid–liquid extraction (37). Optimization of the separation resulted in recovery of product at 93% purity and 95% yield. This type of separation differs markedly from protein purification in feed concentration (≈ 50 –200 g/L for cefonicid vs 1 to 10 g/L for protein), molecular weight of impurities (<5000 compared to 10,000–100,000 for proteins), and throughputs (≈ 1 –2 mg (g stationary phase/min) compared to 0.01–0.1 mg (g/min) for proteins).

Reversed-phase separation was also found to purify diastereomer precursors used in the chemical synthesis of the insect sex pheromone of *Limantria dispar*, a pest which attacks oak trees. The liquid chromatography columns tested had dimensions of up to 15 cm id by 130 cm long, and were able to purify up to 708 g of starting material in 4.1 L sample using a column having 23 L of stationary phase. The throughput is estimated to have been on the order of 0.2–0.4 mg (g/min), where separation was obtained using a gradient of hexane and diethyl ether.

Small Particle Silica Columns. Process-scale reversed-phase supports can have particle sizes as small as 5–25 μm . Unlike polymeric reversed-phase sorbents, these small-particle silica-based reversed-phase supports require high pressure equipment to be properly packed and operated. The introduction of axial compression columns has helped promote the use of high performance silica supports on a process scale. Resolution approaching that of an analytical-scale separation can be achieved using these columns that can also be quickly packed. These columns consist of a plunger fitted into a stainless steel column. The particles are placed into the column in a slurry. The plunger then squeezes or compacts the bed in an axial direction to give a stable, tightly packed bed. This type of column must be operated at pressures of up to 10 MPa (100 bar), but also gives excellent resolution in run times of an hour or less (36).

Hydrophobic Interaction Chromatography. Hydrophobic interactions of solutes with a stationary phase result in their adsorption on neutral or mildly hydrophobic stationary phases. The solutes are adsorbed at a high salt concentration, and then desorbed in order of increasing surface hydrophobicity, in a decreasing kosmotrope gradient. This characteristic follows the order of the lyotropic series for the anions: phosphates > sulfates > acetates > chlorides > nitrates > thiocyanates. Anions which precipitate proteins less effectively than chloride (nitrates and thiocyanates) are chaotropes or water structure breakers, and have a randomizing effect on water's structure; the anions preceding chlorides, ie, phosphates,

sulfates, and acetates, are polar kosmotropes or water structure makers. These promote precipitation of proteins. Kosmotropes also promote adsorption of proteins and other solutes onto a hydrophobic stationary phase (46). These kosmotropes have other beneficial characteristics which include increasing the thermal stability of enzymes, decreasing enzyme inactivation, protecting against proteolysis, increasing the association of protein subunits, and increasing the refolding rate of denatured proteins. Hence, utilization of hydrophobic interaction chromatography is attractive for purification of proteins where recovery of a purified protein in an active and stable conformation is desired (46,47).

Salt Effects. The definition of a capacity factor k' in hydrophobic interaction chromatography is analogous to the distribution coefficient, K_{av} , in gel permeation chromatography:

$$k' = \frac{V_e - V_o}{V_o} \tag{5}$$

However, because protein retention owing to adsorption can occur, the value of k' can be greater than one, ie, elution of the most retained peak need not occur after one column volume of mobile phase has passed through the column. The retention behavior of lysozyme on a polymeric hydrophobic interaction support follows the preferential interaction parameter of the lyotropic series of anions (Fig. 13). The preferential interaction parameter is a measure of the net salt inclusion or exclusion of the hydration layer. The higher the value, the larger the disrupting effect of the salt. This analysis led to derivation and experimental validation of the capacity factor for lysozyme with respect the lyotropic number of the anion for a hydrophilic vinyl polymer support having an average particle diameter of 30 μm , and average pore size of 100 nm. This capacity factor has the following form:

$$K' = a [C]^d [N_x - b]^{-h} \tag{6}$$

where a , b , d , and h are protein specific parameters, N_x is the lyotropic number, and C is salt concentration in M . Hydrophobic interaction parameters can then be estimated for experimental peak retention data as changes in retention time upon a change in salt or salt concentration. An example of how salt type and concentration affect retention of lysozyme is illustrated in Figure 14. A similar functional relation was found for myoglobin with respect to a hydrophilic vinyl polymer derivatized using butyl groups (47).

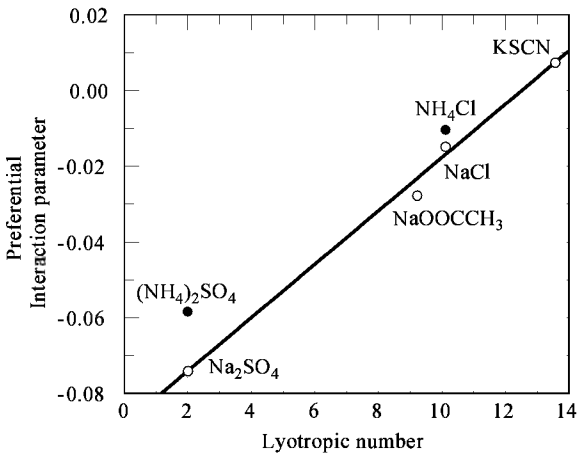


Fig. 13. Preferential interaction parameter vs lyotropic number for lysozyme on (○) bovine serum albumin and (●) Toyopearl.

Courtesy of American Chemical Society (47).

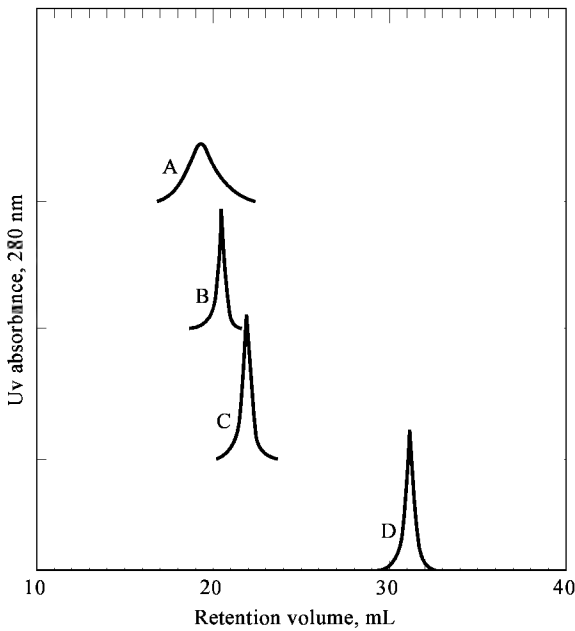


Fig. 14. Chromatographic retention of 20 μL of a 3 mg/mL solution of lysozyme on Toyopearl HW-65S using a 50 cm $\times$ 8 min ID column in 1.3 M ammonium salt, 20 mM Tris mobile phase at 1 mL/min for A, NH_4I ; B, NH_4Cl ; C, $\text{NH}_4\text{OOCCH}_3$; and D, $(\text{NH}_4)_2\text{SO}_4$.

Courtesy of American Chemical Society (47).

Various types of proteins have been purified using hydrophobic interaction chromatography including alkaline phosphatase, estrogen receptors, isolectins, strepavidin, calmodulin, epoxide hydrolase, proteoglycans, hemoglobins, and snake venom toxins (46). In the case of cobra venom toxins, the order of elution of the six cardiotoxins supports the hypothesis that the mechanism of action is related to hydrophobic interactions with the phospholipids in the membrane.

The recovery of recombinant chymosin from a yeast fermentation broth showed that large-scale hydrophobic interaction chromatography could

produce an acceptable product in one step. Chymosin, which used to be obtained from the lining of the stomachs of calves, is used in cheesemaking, and its cost is an issue. Because the capacity of the hydrophobic interaction stationary-phase is limited, an alternative method has been developed in which the enzyme is extracted into a two-phase polyethylene glycol (PEG) salt system. The partition for the chymosin into PEG coefficient is 100, and hence enables efficient recovery of this in one step. Together with a subsequent ion-exchange step, this method gives a suitably purified chymosin. The use of hydrophobic interaction chromatography may have helped to indicate that two-phase extraction is a viable approach (10).

Affinity Chromatography. The concept of affinity chromatography, credited to the discovery of biospecific adsorption in 1910, was reintroduced as a means to purify enzymes in 1968 (49). Substrates and substrate inhibitors diffuse into the active sites of enzymes irreversibly or reversibly binding there. Conversely, if the substrate or substrate inhibitor is immobilized through a covalent bond to a solid particle of stationary phase having large pores, the enzyme should be able to diffuse into the stationary phase and bind with the substrate or inhibitor. Because the substrate is small (mol wt < 500) and the enzyme large (>15,000), the diffusion of the enzyme to its binding partner at a solid surface can be sterically hindered. The placement of the substrate at the end of an alkyl or glycol chain tethered to the stationary phase's surface reduces hindrance and forms the basis of affinity chromatography. This concept has also been applied to ion-exchange chromatography under the names of tentacle or fimbriated stationary phases.

The realization that enzymes could be selectively retained in a chromatography column packed with particles of immobilized substrates or substrate analogues led to experiments with other pairs of binding partners. Numerous applications of affinity chromatography developed, given the specific and reversible yet strong affinity of biological macromolecules for numerous specific ligands or effectors. These interactions have been exploited for purposes of highly selective, but often expensive protein purifications, recovery of messenger ribonucleic acid (mRNA) in some recombinant DNA applications, and study of mechanisms of protein binding with effector molecules (49).

Minimization of Nonspecific Binding. The purpose of affinity chromatography is the highly selective adsorption and subsequent recovery of the target biomolecule. Loss of specificity occurs when macromolecules, other than the targeted materials, adsorb onto the stationary phase owing to hydrophobic or ionic interactions. For example, a spacer arm, which allows the binding ligand on the column to be located away from the matrix surface, can improve accessibility and reduce steric hindrance to the immobilized ligand, often decreasing selectivity. Hexamethylenediamine, a common spacer arm used initially in affinity chromatography, has been the source of strong, nonspecific binding. This hydrophobic character has been decreased by interposing an ether or secondary amine, such as 3,3-diaminodipropylamine, in the middle of the spacer arm.

The ideal matrices for anchoring binding ligands are nonionic, hydrophilic, chemically stable, and physically robust. The most popular matrices are polysaccharide based, principally owing to their hydrophilic character and history of use as size-exclusion or gel-permeation gels (see Fig. 6), although glass beads, polyacrylamide gels, cross-linked dextrans (Sephadex), and agarose synthesized into a bead form have all been used (49). In particular, agarose such as Biogel A (Bio-Rad), Ultragel A (IBF), and Sepharose (Pharmacia) are popular (49,50). Cross-linked agaroses (Sepharose CL and Sepharose FF by Pharmacia) are physically more stable than Sepharose and are suitable for attaching affinity ligands. Both forms of the agaroses have an open porosity which allows proteins to readily diffuse inside. Affinity chromatography results are usually reported in terms of specific activity of the final product (activity per mg of material) and the amount of biomolecule recovered (% yield).

Activation of the Stationary-Phase Surface. Activation of polysaccharide, silica, or polyacrylamide stationary phases involve the formation of a reactive intermediate, covalently attached to the surface, to which a difunctional alkyl-, aryl-, or glycol spacer is subsequently joined. The other end of the spacer is subsequently reacted with the ligand. Cyanogen bromide, CNBr, has been widely used to activate agarose and dextran gels (49). The attachment of ligands, and sometimes activation of supports, is generally carried out in the laboratories of the chromatography process developers because fully prepared affinity stationary phases are not as widely available as stationary phases for the other types of chromatography.

Multistep Processes. An excellent synopsis and industrial viewpoint of affinity chromatography and its fit with other bioseparations unit operations is available (49,50). Ligands range from the low molecular weight components, eg, arginine and benzamidine, which both bind trypsin-like proteases, triazine dyes, and metal chelates; to high molecular weight ligands, eg, protein A, immunoglobins, and monoclonal antibodies. The blood factor VIII, purified by monoclonal affinity techniques, was approved by the U.S. FDA in 1988. Limitations of affinity chromatography as an industrial separation technique can be due to leaching of bound ligands from the column into the product at ppm levels, nonspecific interactions resulting in contamination of the target molecule, and failure of the affinity ligand to differentiate all variant forms of a protein or polypeptide (52). For example, polyclonal antibodies do not distinguish desamido-insulin, which contains a deamidated asparagine, from insulin. Because many antibody preparations cannot differentiate between minor structural changes in proteins, affinity chromatography must be followed by other separation steps, and does not provide a one-step purification.

Receptor Affinity Chromatography. Receptor affinity chromatography is a selective form of immunoaffinity chromatography which is based on antigen-antibody interactions (52). Protein or polypeptide ligands used in preparing receptor affinity supports are themselves products of fermentation of recombinant microorganisms and are subjected to a separate sequence of purification steps, prior to being reacted with a functionalized stationary phase to form the affinity support. The resulting affinity chromatography columns are expensive when viewed on the basis of cost of support unit volume of stationary phase. The cost/benefit ratio would still be attractive because process-scale columns can be small (volumes on the order of 1–10 L). Moreover, as with other types of affinity chromatography, purification of dilute but highly active protein is possible.

BIBLIOGRAPHY

1.

R. C. Willson and M. R. Ladisch, in M. R. Ladisch, R. C. Willson, C-C. Painton, and S. E. Builder, eds., *ACS Symp. Ser.* **427**, 1–13 (1990).

2.

Committee on Bioprocess Engineering, National Research Council, *Putting Biotechnology to Work: Bioprocess Engineering*, National Academy of Sciences, Washington, D.C., 1992, pp. 2–22.

3.

S. M. Wheelwright, *Protein Purification: Design and Scale-up of Downstream Processing*, Hanser Publishers, Munich, Germany, 1991, pp. 1–9, 61, 213–217.

4.

C. A. Bisbee, *GEN* **13**(14), 8–9 (1993).

5.

A. M. Thayer, *C&EN* **74**(33), 13–21 (1996).

6.

A. M. Thayer, *C&EN* **74**(1), 22–23 (1996).

7.

R. Rawls, *C&EN* **74**(32), 31 (1996).

8.

J. A. Wells, *Science* **273**(5274), 449–450 (1996).

9.

W. Roush, *Science* **273**(5273), 300–301 (1996).

10.

V. B. Lawlis and H. Heinsohn, *LC-GC* **11**(10), 720–729 (1993).

11.

M. Bernon and J. Bodelle, in Y-Y. H. Chien and J. L. Gueriguian, eds., *Drug Biotechnol. Reg.* **13**, xv–xxiii (1991).

12.

M. R. Ladisch and K. L. Kohlmann, *Biotechnol. Prog.* **8**(6), 469–478 (1992).

13.

W. F. Prouty, in Ref. 11, pp. 221–262.

14.

E. P. Kroeff, R. A. Owens, E. L. Campbell, R. D. Johnson, and H. I. Marks, *J. Chromatog.* **461**, 45–61 (1989).

15.

S. E. Builder, R. van Reis, N. Paoni, and J. Ogez, "Process Development and Regulatory Approval of Tissue-Type Plasminogen Activator," in *Proceedings of the 8th International Biotechnology Symposium*, Paris, July 17–22, 1989.

16.

N. K. Harakas, J. P. Schaumann, B. T. Connolly, A. J. Wittwer, J. V. Orlander, and J. Feder, *Biotechnol. Prog.* **4**(3), 149–158 (1988).

17.

S. E. Builder and E. Grossbard, *Transfus. Med. Rec. Technol. Adv.*, 303–313 (1986).

18.

M. Barinaga, *Science* **272**(5262), 664–666 (1996).

19.

J. O'C. Hamilton, *Business Week*, **3478**, 118, 122 (1996).

20.

S. E. Builder and W. S. Hancock, *Chem. Eng. Prog.* **84**(8), 42–46 (1988).

21.

J. D. Watson, M. Gilman, J. Witkowski, and M. Zoller, *Recombinant DNA*, 2nd ed., W. H. Freeman and Co., New York, 1992, pp. 458–460.

22.

B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 2nd ed., Garland Publishing, New York, 1989, pp. 1203–1212.

23.

S. V. Ho, in Ref. 1, pp. 14–34.

24. G. B. Dove, G. Mitra, G. Roldan, M. A. Shearer, and M-S. Cho, in Ref. 1, pp. 194–209.

25. S. Sobel, in Ref. 11, pp. 499–511.

26. J. K. Lin, B. J. Jacobsen, A. N. Pereira, and M. R. Ladisch, in W. A. Wood and S. T. Kellog, eds., *Methods in Enzymology*, Vol. 160, Academic Press, Inc., San Diego, Calif., 1988, pp. 145–159.

27. A. Velayudhan and M. R. Ladisch, *Anal. Chem.* **63**(18), 2028–2032 (1991).

28. A. Velayudhan, R. L. Hendrickson, and M. R. Ladisch, *AIChE J.* **41**(5), 1184–1193 (1995).

29. A. Velayudhan and M. R. Ladisch, *Ind. Eng. Chem. Res.* **34**(8), 2805–2810 (1995).

30. S. Roe, in E. L. V. Harris and S. Angal, eds., *Protein Purification Methods—A Practical Approach*, IRL Press, Oxford, U.K., 1989, pp. 200–216.

31. A. Z. Preneta, in Ref. 30, pp. 293–306.

32. R. T. Kurnik, A. W. Yu, G. S. Blank, A. R. Burton, D. Smith, A. M. Athalye, and R. van Reis, *Biotechnol. Bioeng.* **45**, 149–157 (1995).

33. K. H. Hamaker, J. Liu, R. J. Seely, C. M. Ladisch, and M. R. Ladisch, *Biotechnol. Prog.* **12**, 184–189 (1996).

34. *Gel Filtration—Theory and Practice*, Pharmacia Fine Chemicals, Uppsala, Sweden, 1979, pp. 1–19, 30–35, 46, 50, and 57; *Gel Filtration—Principles and Methods*, 6th ed., Pharmacia Biotech, 1997.

35. P. C. Sadek, P. W. Carr, R. M. Doherty, M. J. Kamlet, R. W. Taft, and M. H. Abraham, *Anal. Chem.* **57**, 2971–2978 (1985).

36. H. Colin, P. Hilaireau, and J. De Tournemire, *LC-GC* **8**(4), 302–312 (1990).

37. A. M. Cantwell, R. Calderone, and M. Sienko, *J. Chromatog.* **316**, 133–149 (1984).

38. R. L. Gustafson, R. L. Albright, I. Heisler, J. A. Lirio, and O. T. Reid, *Ind. Eng. Chem. Prod. Res. Dev.* **7**(2), 107–115 (1968).

39. D. J. Pietrzyk and J. D. Stodola, *Anal. Chem.* **53**(12), 1822–1828 (1981).

40. D. J. Pietrzyk and C-H. Chu, *Anal. Chem.* **49**(6), 757–764 (1977).

41. J. Morris and J. S. Fritz, *LC-GC* **11**(7), 513–517 (1993).

42. G. K. Sofer and L. E. Nystrom, *Process Chromatography: A Practical Guide*, Academic Press, Ltd., London, 1989, pp. 128–129.

43. R. L. Garnick, M. J. Ross, and R. A. Baffi, in Ref. 11, pp. 263–313.

44. E. A. Carrey, in T. E. Creighton, ed., *Peptide Mapping in Protein Structure: A Practical Approach*, IRL Press, Oxford, U.K., 1990, pp. 117–144.

45. E. Flanigan, *High Performance Liquid Chromatography in the Production and Quality Control of Salmon Calcitonin*, Purdue University, West Lafayette, Ind., 1991, p. 207.

46. B. F. Roettger and M. R. Ladisch, *Biotechnol. Adv.* **7**, 15–29 (1989).

47. B. R. Roettger, J. A. Myers, M. R. Ladisch, and F. E. Regnier, in Ref. 1, pp. 80–92.

48. M. R. Ladisch, R. L. Hendrickson, and K. L. Kohlmann, in Ref. 1, pp. 93–103.

49. I. Parikh and P. Cuatrecases, *CEN* **63**, 17–32 (1985).

50. J. A. Asenjo and I. Patrick, in E. L. V. Harris and S. Angal, eds., *Protein Purification Applications—A Practical Approach*, IRL Press, Oxford, U.K., 1990, pp. 1–28.

51. L. Hagel, "Gel Filtration" in J-C. Janson and L. Ryden, ed., *Protein Purification: Principles, High Resolution Methods, and Applications*, VCH Publishers, N.Y., 1989, pp. 63–106.

52. P. Bailon and D. V. Weber, *Nature* **335**(6193), 839–840 (1988).

53. S. K. Basak and M. R. Ladisch, *Anal. Biochem.* **226**, 51–58 (1995).

Michael R. Ladisch
Purdue University

CHIRAL SEPARATIONS

Chiral separations are concerned with separating molecules that can exist as nonsuperimposable mirror images. Examples of these types of molecules, called *enantiomers* or *optical isomers*, are illustrated in Figure 1. Although chirality is often associated with compounds containing a tetrahedral carbon with four different substituents, other atoms, such as phosphorus or sulfur, may also be chiral. In addition, molecules containing a center of asymmetry, such as hexahelicene, tetrasubstituted adamantanes, and substituted allenes or molecules with hindered rotation, such as some 2,2' disubstituted binaphthyls, may also be chiral. Compounds exhibiting a center of asymmetry are called *atropisomers*. An extensive review of stereochemistry may be found under PHARMACEUTICALS, CHIRAL.

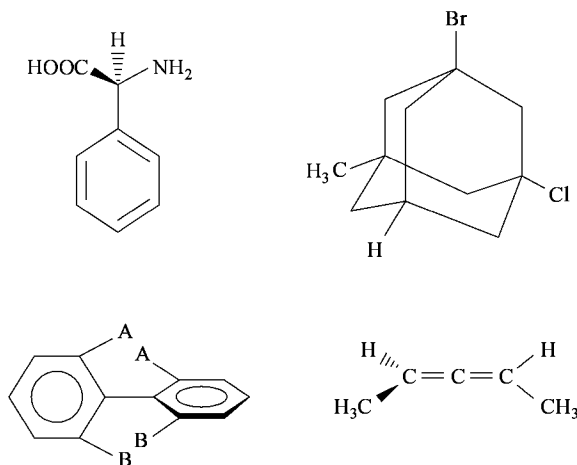
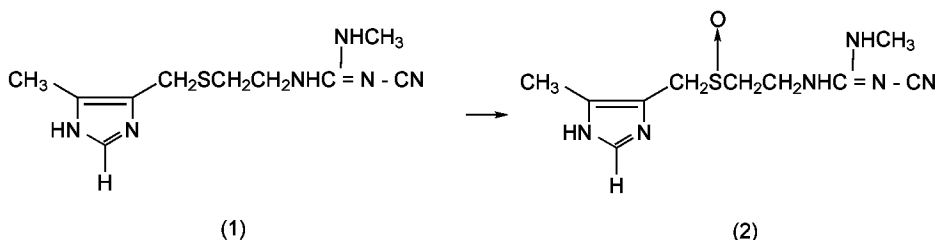


Fig. 1. Examples of chiral molecules.

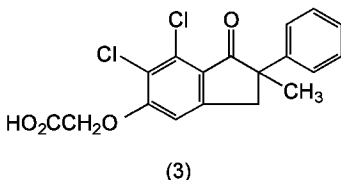
Although scientists have known since the time of Louis Pasteur (1) that optical isomers can behave differently in a chiral environment (eg, in the presence of polarized light), it has only been since about 1980 that there has been a growing awareness of the implications arising from the fact that many drugs are chiral and that living systems constitute chiral environments. Hence, the optical isomers of chiral drugs may exhibit different bioactivities and or biotoxicities.

In the case of enantiomerically pure chiral drugs, the possibility of racemization or inversion either *in vivo* or during storage cannot be ruled out. Ibuprofen is an example of a chiral drug which undergoes rapid inversion *in vivo* (2). In addition, there are several examples of achiral (or *prochiral*) drugs being biotransformed into chiral entities. In some cases, the enantiomeric ratios produced by laboratory animals may differ from that produced in humans. For example, cimetidine (1), used to treat peptic ulcers and marketed as Tagamet, is achiral. However, cimetidine sulfoxide (2), one of its major metabolites, is chiral by virtue of oxidation of the sulfur atom to a sulfoxide (the lone pair of electrons on the sulfur constitutes the fourth group). In humans, the (+) enantiomer predominates (2.4:1) but in rats, although (+)-cimetidine sulfoxide is produced in excess, the enantiomeric ratio approaches racemic (1.3:1). This raises the question of the suitability of rats as appropriate test models for this particular drug (3).



For those drugs that are administered as the racemate, each enantiomer needs to be monitored separately yet simultaneously, since metabolism, excretion or clearance may be radically different for the two enantiomers. Further complicating drug profiles for chiral drugs is that often the pharmacodynamics and pharmacokinetics of the racemic drug is not just the sum of the profiles of the individual enantiomers.

Although it might seem that administration of enantiomerically pure substances would always be preferred, the diuretic indacinone (3), is an example of a drug for which one enantiomer mediates the harmful effects of the other enantiomer (4). (+)-Indacinone, the diuretically active enantiomer or *eutomer*, causes uric acid retention. Fortunately, the other enantiomer (*distomer*) causes uric acid elimination. Thus, administration of a mixture of the two enantiomers, although not necessarily racemic, may have therapeutic value.



Although a great deal of the work currently being done in chiral separations is related to pharmaceuticals, the agricultural and the food and beverage industries are affected as well. For instance, several chiral pesticides are used commercially. It is possible that the enantiomers may differ in their persistence in the environment and their effectiveness against specific pests. For example, the neurotoxic action of the pesticide, ethyl-4-nitrophenyl phenylphosphono thionate (EPN), resides almost entirely in the *S* enantiomer while the desired insecticidal activity resides entirely in the *R* enantiomer (5). This raises the question of whether the pesticide may be safer and more effective if applied as an enantiomerically pure formulation. In the food and beverage industry,

many of the constituents that confer flavor or aroma in foods and beverages are chiral. For instance, the configuration of the 4-alkyl-substituted γ -lactones responsible for much of the flavor in fruits is almost exclusively R (6). Often, the two enantiomers have very different aromas or flavors. The presence of any of the "unnatural" enantiomer may confer an "off-flavor" to the substance and may be indicative of racemization under adverse storage conditions, adulteration, or formulation from nonnatural sources.

The growing awareness of the implications of chirality to the pharmaceutical industry has spurred tremendous effort toward stereoselective synthetic strategies and the development of new chiral catalysts. However, the enantiomeric purity of these substances or their chiral precursors needs to be determined. Also, there are many chiral compounds for which no stereospecific synthetic pathways have been devised. Thus, there is a tremendous need not only for analytical scale (<5–10 mg), but bulk-scale chiral separations as well. Whether analyzing drugs or synthetic precursors for enantiomeric purity, monitoring biological or environmental samples for chiral discrimination or trying to enantioresolve kilogram quantities of a racemic drug, there are a variety of reasons for performing chiral separations. The purpose of the separation dictates, to some extent, the method employed.

Traditionally, chiral separations have been considered among the most difficult of all separations. Conventional separation techniques, such as distillation, liquid–liquid extraction, or even some forms of chromatography, are usually based on differences in analyte solubilities or vapor pressures. However, in an achiral environment, enantiomers or optical isomers have identical physical and chemical properties. The general approach, then, is to create a "chiral environment" to achieve the desired chiral separation and requires chiral analyte–chiral selector interactions with more specificity than is obtainable with conventional techniques.

A variety of strategies have been devised to obtain chiral separations. Although the focus of this article is on chromatographically based chiral separations, other methods include crystallization and stereospecific enzymatic-catalyzed synthesis or degradation. In crystallization methods, racemic chiral ions are typically resolved by the addition of an optically pure counterion, thus forming diastereomeric complexes.

Enzymatically based methods depend on the stereospecificity of an enzyme-catalyzed reaction, such as lipase-catalyzed esterification, to degrade enantioselectively the unwanted enantiomer or to produce the desired enantiomer. Because only one enantiomer undergoes the reaction, the subsequent separation is reduced to separating two different species. For example, in the case of enzyme-catalyzed esterification, the originally difficult enantiomeric separation is reduced to the separation of the ester of one optical isomer from the alcohol or acid of the other optical isomer of the original starting material, and may be accomplished using a variety of conventional separation methodologies (7). One disadvantage of enzymatically based methods is that only one enantiomer is obtained and there is usually no analogous method for producing the opposite enantiomer.

An alternative method of creating a chiral environment is to derivatize a chiral analyte with an optically pure reagent, thus, producing diastereomers. The resultant diastereomers, containing more than one chiral center, have slightly different melting and boiling points and can often be separated using conventional methods. A number of chiral derivatizing agents, as well as the types of compounds for which they are useful, have been developed and are listed in Table 1. Limitations of this approach include lack of suitable functionality in the analyte that can be derivatized with an appropriate enantiomerically pure derivatizing agent, unavailability of a suitable derivatizing agent of sufficiently high or at least known optical purity, difficulty of removing the derivatizing group after the desired separation has been accomplished, enantiodiscrimination during derivatization, potential racemization either during derivatization or removal or the chiral derivatizing group (which is not always possible), and the additional validation required to confirm that the enantiomeric ratio of the final product corresponds to the original enantiomeric ratio.

Table 1. Analyte Functional Groups and Chiral Derivatizing Reagents

Analyte functional group	Derivatizing agent	Product	Examples of derivatizing agents
carboxylic acid (acid or base catalyzed)	alcohol	ester	(–)-menthol
	amine	amide	1-phenylethylamine
amine (1°) amine (1° and 2°)	aldehyde	isoindole	1-(1-naphthyl)ethylamine
			<i>o</i> -phthaldialdehyde–2-mercaptoethanol
	anhydrides	amide	γ -butyloxycarbonyl-L-leucine anhydride
			<i>O,O</i> -dibenzoyltartaric anhydride
	acyl halides	amide	(R)-(-)-methylmandelic acid chloride
			α -methoxy- α -trifluoromethylphenylacetyl chloride
(1°, 2°; can <i>N</i> -dealkylate 3°)	isocyanates	urea	α -methylbenzyl isocyanate
			1-(1-naphthyl)ethyl isocyanate
	isothiocyanate	thiourea	2,3,4,6-tetra- <i>O</i> -acetyl- β -D-glucopyranosyl isothiocyanate
			α -methylbenzyl isothiocyanate
	chloroformates	carbamate	(–)-menthyl chloroformate
			(+)-1-(9-fluorenyl)ethylchloroformate
alcohols	acyl halides	ester	(–)-menthoxy acid chloride
			(<i>S</i>)- <i>O</i> -propionylmandelyl chloride
	anhydrides	ester	(<i>S,S</i>)-tartaric anhydride
	chloroformate	carbonate	(–)-menthyl chloroformate
	isocyanate	carbamate	α -methylbenzyl isocyanate

Use of Chiral Additives

Another method for creating a chiral environment is to add an optically pure chiral selector to a bulk liquid phase. Historically, many of the chiral selectors currently available as chiral stationary phases for high performance liquid chromatography originated as chiral mobile-phase additives, particularly in thin-layer chromatography (tlc). Chiral additives have several advantages over chiral stationary phases and continue to be the predominant mode for chiral separations by tlc (8) and capillary electrophoresis (ce) (9). First of all, the chiral selector added to a bulk liquid phase can be readily changed. The use of chiral additives allows chiral separations to be done using less expensive, conventional stationary phases. A wider variety of chiral selectors are available to be used as chiral additives than are available as chiral stationary phases, thus, providing the analyst with considerable flexibility. Finally, the use of chiral additives may provide valuable insight into the chromatographic conditions and or likelihood of success with a potential chiral stationary-phase chiral selector. This is particularly important for the development of new chiral stationary phases because of the difficulty and cost involved.

Chiral additives, however, do pose some unique problems. Many chiral agents are expensive or are not commercially available, and therefore, must be synthesized. The presence of the chiral additive in the bulk liquid phase may also interfere with detection or recovery of the analytes. Finally, the presence of enantiomeric impurity in the chiral additive may add analytical complications (10).

Thin-Layer Chromatography. Thin-layer chromatography (tlc) offers several advantages for chiral separations and in the development of new chiral stationary phases. Besides being inexpensive, tlc can be used to screen mobile-phase conditions rapidly (ie, organic modifier content, pH, etc), chiral selectors, and analytes. Several different analytes may be run simultaneously on the same plate. Usually, no preequilibration of the mobile phase and stationary phase is required. In addition, only small amounts of mobile phase, and therefore, chiral mobile-phase additive, are required. Another significant advantage is that the analyte can always be unambiguously found on the tlc plate.

Two mechanisms for chiral separations using chiral mobile-phase additives, analogous to models developed for ion-pair chromatography, have been

proposed to explain the chiral selectivity obtained using chiral mobile-phase additives. In one model, the chiral mobile-phase additive and the analyte enantiomers form "diastereomeric complexes" in solution. As noted previously, diastereomers may have slightly different physical properties such as mobile phase solubilities or slightly different affinities for the stationary phase. Thus, the chiral separation can be achieved with conventional columns.

An alternative model has been proposed in which the chiral mobile-phase additive is thought to modify the conventional, achiral stationary phase *in situ*, thus, dynamically generating a chiral stationary phase. In this case, the enantioseparation is governed by the differences in the association between the enantiomers and the chiral selector in the stationary phase.

Several different types of chiral additives have been used including (1R)-(-)-ammonium-10-camphorsulfonic acid (11), cyclodextrins (12,13), proteins, and various amino acid derivatives such as *N*-benzoxycarbonyl-glycyl-L-proline as well as macrocyclic antibiotics (14). Chiral counterions such as (1R)-(-)-ammonium-10-camphorsulfonic acid and *N*-benzoxycarbonyl-glycyl-L-proline have been used under normal phase conditions (eg, ca 2.5 mM with 1 mM triethylamine in methylene chloride on a diol tlc plate), promoting ion-pair associations (15). In contrast, the cyclodextrins (16,17), proteins, and amino acid derivatives have been used exclusively under aqueous mobile phase conditions. In the case of the cyclodextrins, the limited aqueous solubility of β -cyclodextrin ($\sim 0.17\text{ M}$ at room temperature), the most commonly used cyclodextrin, can be enhanced by using a saturated urea solution. In addition, it is recommended that 0.6 M NaCl can be used to stabilize the binder by which the stationary phase is attached to the glass support (18).

Chiral separation validation in tlc may be accomplished by recovering the individual analyte spots from the plate and subjecting them to some type of chiroptical spectroscopy such as circular dichroism or optical rotary dispersion. Alternatively, the plates may be analyzed using a scanning densitometer. Scanning densitometers irradiate the surface of the plate at a specified wavelength (in the ultraviolet or visible regions) and can measure the intensity of the reflected beam. A trace of the reflected beam vs distance has the general appearance of a chromatogram and an example is shown in Figure 2 (19). The relative peak heights or areas of the two enantiomers obtained at two or more different wavelengths should remain constant because the extinction coefficients of the enantiomers are identical at every wavelength.

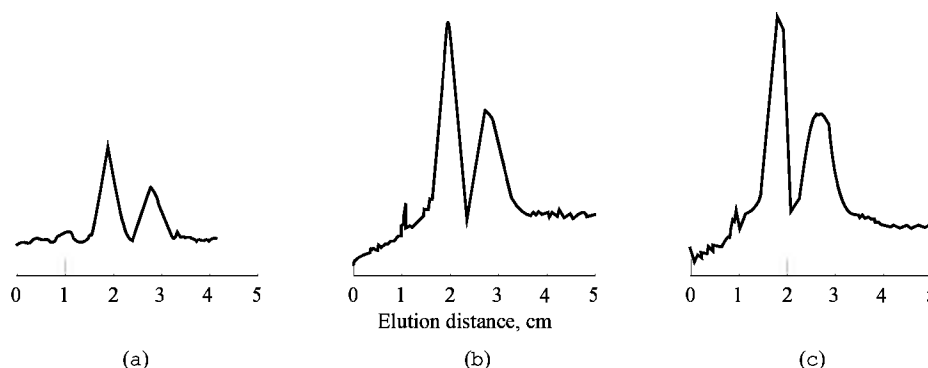


Fig. 2. Tlc densitometer scans showing the resolution of isoproterenol on a hplc silica-gel plate obtained using a mobile phase consisting of 6.8 mM (1R)-(-)-ammonium-10-camphorsulfonic acid in 75:25 (v/v) methylene chloride:methanol. (a) 254 nm, (b) 275 nm, (c) 300 nm.

Capillary Electrophoresis. Capillary electrophoresis (ce) or capillary zone electrophoresis (cze), a relatively recent addition to the arsenal of analytical techniques (20,21), has also been demonstrated as a powerful chiral separation method. Its high resolution capability and lower sample loading relative to hplc makes it ideal for the separation of minute amounts of components in complex biological mixtures (22,23).

In a ce experiment, a thin capillary is filled with a run buffer and a voltage is applied across the capillary. Although a complete treatment of the fundamental principles of ce is beyond the scope of this article, it can be said that the underlying impetus for separations in ce is, in general, derived from the fact that charged species migrate in response to an applied electric field proportionately to their charge and inversely proportionately to their size. Thus, given equivalent charges, lighter analytes have higher electrophoretic mobilities than heavier analytes and, given equivalent sizes, more highly charged species have higher mobilities than lesser charged or neutral species. In fact, neutral species have no intrinsic electrophoretic mobility. Species having opposite charges have electrophoretic mobilities in opposing directions.

Chiral separations by ce have been performed almost exclusively using chiral additives to the run buffer. The advantages of this approach are identical to the advantages mentioned previously with regard to using chiral mobile-phase additives in tlc. Many of the chiral selectors used successfully as mobile-phase additives in tlc and as immobilized ligands in hplc have been used successfully in ce including proteins (24), native (25) and functionalized cyclodextrins (26,27), various carbohydrates (28,29), assorted functionalized amino acids (30), chiral-ion pairing agents (31), and macrocyclic antibiotics (32). Other ce chiral selectors which have not been used as immobilized chiral selectors in hplc include bile salts (33), chiral surfactants (34), and dextran sulfate (35).

Although chiral ce is most commonly performed using aqueous buffers, there has been some work using organic solvents such as methanol, formamide, *N*-methylformamide or *N,N*-dimethylformamide with chiral additives such as quinine (36) or cyclodextrins (37,38). Nonaqueous ce requires that the background electrolyte be prepared using organic acids (eg, citric acid or acetic acid) and organic bases (eg, tetraalkylammonium halides or tris(hydroxymethyl)-aminomethane).

Theoretical models (39,40), as expressed in equation 1, where μ represents the mobility

$$\mu_1 - \mu_2 = \frac{[\mu_{1,c} - \mu_{1,f}] [K_1 - K_2] [CA]}{[1 + K_1[CA]] [1 + K_2[CA]]} \quad (1)$$

of the analyte in the free and complexed states, K represents the binding constants of enantiomers 1 and 2 and $[CA]$ is the molar concentration of the additive, reveal that, in general, two conditions must be met to achieve chiral separations by ce. First, there must be differences in the binding constants of the two enantiomers with the chiral selector. Second, because the intrinsic electrophoretic mobilities of the enantiomers are identical in the free state, there must be a significant difference in the mobilities of the analyte in the complexed and free state. Chiral selectors have generally been shown to be the most effective when the intrinsic electrophoretic mobility of the additive is in the opposite direction of the intrinsic electrophoretic mobility of the analyte.

Chiral separations by ce is a rapidly growing field and offers the analyst tremendous flexibility with regard to chiral selector choice. In addition, because the additive is typically in the run buffer, there is virtually no column preequilibration. However, ce instruments tend to cost more than most chromatographic systems, sample capacity is much smaller for ce than for an analogous hplc method, sample recovery is not a trivial problem in ce and run-to-run reproducibility for ce tends to be much worse than for most chromatographic methods. Nevertheless, the flexibility, minimal sample and/or chiral selector required and the extremely high resolving power of ce ensure that this technique will continue to play an important role in chiral separations in the future.

Chiral Stationary Phases

Most chiral chromatographic separations are accomplished using chromatographic stationary phases that incorporate a chiral selector. The chiral separation mechanisms are generally thought to involve the formation of transient diastereomeric complexes between the enantiomers and the stationary phase chiral ligand. Differences in the stabilities of these complexes account for the differences in the retention observed for the two enantiomers. Often, the use of a

chiral stationary phase allows for the direct separation of the enantiomers without the need for derivatization. One advantage offered by the use of chiral stationary phases is that the chiral selector need not be enantiomerically pure, only enriched. In addition, for chiral stationary phases having a well understood chiral recognition mechanism, assignment of configuration (eg, R or S) may be possible even in the absence of optically pure standards. However, chiral stationary phases have some limitations. The specificity required for chiral discrimination limits the broad applicability of most chiral stationary phases; thus there is no "universal" chiral stationary phase. The cost of most chiral columns are typically much higher (~3×) than for conventional columns. In contrast to conventional chromatographic columns, chiral stationary phases are generally not as robust, require more careful handling than conventional columns and usually, once column performance has begun to deteriorate, cannot be returned to their original performance levels. In many cases, chromatographic column choice or mobile-phase optimization for chiral stationary phases is not as straightforward as with conventional stationary phases. In conventional chromatography, there is usually a well-behaved relationship between retention and mobile phase composition or column temperature. However, in many of the chiral stationary phases, the stationary phase present a multitude of different types of sites with not necessarily equivalent populations for interaction with the analytes. In the case of some liquid chromatographic stationary phases, the different types of sites may result in normal phase type behavior under very nonpolar mobile phase conditions and reversed-phase type behavior under highly polar mobile phase conditions. The multiplicity of types and numbers of sites also confounds thermodynamic considerations (41). Often, there is a narrow window of mobile-phase conditions under which enantioselectivity is observed and these conditions may be unique for a particular chiral analyte. Thus, for many of the chiral stationary phases, adequate chiral recognition models, used to guide selection of the appropriate column for a given separation, have yet to be developed. Column selection, therefore, is often reduced to identifying structurally similar analytes for which chiral resolution methods have been reported in the scientific literature or chromatographic supply catalogues and adapting a reported method for the chiral pair to be resolved.

An additional complication, sometimes arising with the use of chiral stationary phases, may occur when the analytes either exist as *conformers* or can undergo inversion during the chromatographic analysis. Figure 3 illustrates a typical chromatogram obtained for oxazepam, one of the chiral benzodiazepines that can undergo ring opening and inversion at the chiral center (42). As can be seen from Figure 3, the peaks appear to have a plateau between them and are sometimes referred to as "Batman peaks". This effect can sometimes be suppressed by lowering the column temperature. Although the appearance of Batman peaks is not unique to chiral separations, the specificity of chiral analyte–chiral selector interactions may increase the frequency of their occurrence.

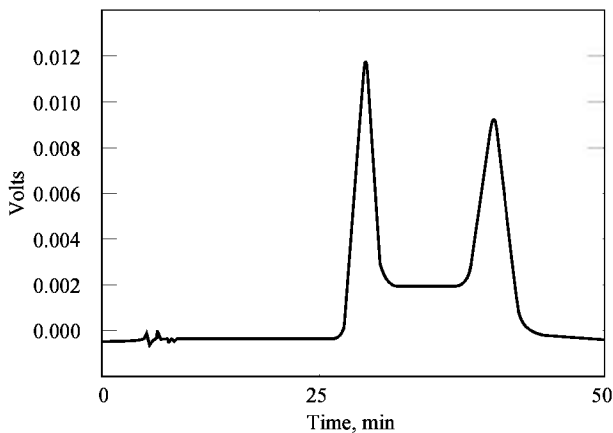


Fig. 3. The chiral separation obtained for oxazepam on a sulfated cyclodextrin hplc column (4.6 mm ID × 25 cm) using a 10% acetonitrile buffer (25 mM ammonium acetate, pH 7).

Thin-Layer Chromatography. Chiral stationary phases have been used less extensively in tlc as in high performance liquid chromatography (hplc). This may, in large part, be due to lack of availability. The cost of many chiral selectors, as well as the accessibility and success of chiral additives, may have inhibited widespread commercialization. Usually, nondestructive visualization of the sample spots in tlc is accomplished using iodine vapor, uv or fluorescence. However, the presence of the chiral selector in the stationary phase can mask the analyte and interfere with detection (43).

Chiral stationary phases in tlc have been primarily limited to phases based on normal or microcrystalline cellulose (44,45), triacetylcellulose sorbents or silica-based sorbents that have been chemically modified (46) or physically coated to incorporate chiral selectors such as amino acids (47,48) or macrocyclic antibiotics (49) into the stationary phase.

Of the silica-based materials, only the ligand-exchange phases are commercially available (Chiralplate, tlc plates are available through Alltech Associates, Inc.) Supelco, Inc., the Aldrich Chemical Company, and Bodman Industries are all based on ligand exchange. Typically in the case of the ligand-exchange type tlc plates, the ligand-exchange selector is comprised of an amino acid residue to which a long hydrocarbon chain has been attached (eg, (2S,4R,2'RS)-4-hydroxy-1-(2-hydroxydodecyl)proline) (50). The hydrocarbon chain of the functionalized amino acid is either chemically bonded to the substrate or intercalates in between the chains of a reversed phase-stationary phase thus immobilizing the chiral selector. The bidentate amino acid chiral selector is thought to reside close to the surface of the stationary phase and participates as a ligand in the formation of a bi-ligand complex with a divalent metal ion (eg, Cu²⁺) and the chiral bidentate analyte (Fig. 4). Analytes enantioselectable using ligand exchange are usually restricted to 1,2-diols, α-amino acids, α-amino alcohols, and α-hydroxyacids (51,52). Again, differences in the stabilities of the diastereomeric complexes thus formed give rise to the chiral separation.

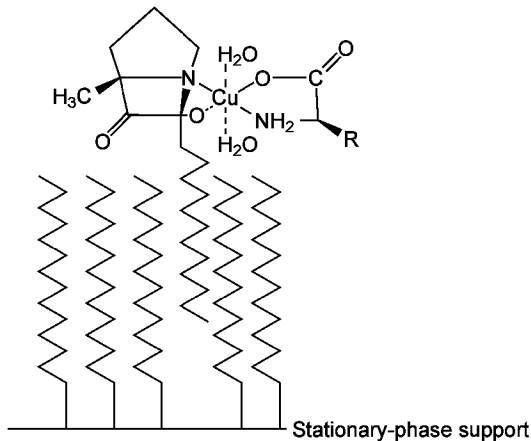


Fig. 4. A ligand-exchange chiral selector complexed with a chiral analyte.

High Performance Liquid Chromatography. Although chiral mobile phase additives have been used in high performance liquid chromatography (hplc), the large amounts of solvent, thus chiral mobile phase additive, required to pre-equilibrate the stationary phase renders this approach much less attractive than for tlc and is not discussed here.

The last decade has seen the commercialization of a large number of different types of chiral stationary phases including the cyclodextrin phases (53), the chirobiotic phases (54), the π - π interaction phases (55,56), the protein phases (57–61), as well as the cellulosic and amylosic phases (62,63) and chiral crown ether phases (64,65). Currently, there are over 50 different chiral columns that are commercially available for hplc. Table 2 briefly summarizes the types of columns available as well as typical applications and mobile-phase conditions. Each of these chiral stationary phases are very successful at separating large numbers of enantiomers, which in many cases, are unresolvable using any of the other chiral stationary phases. Unfortunately, despite the large number and variety of chiral stationary phases currently available, there remains a large number of enantiomeric compounds that are unresolvable by any of the existing chiral stationary phases. In addition, incomplete understanding of the chiral recognition mechanisms of many of these chiral stationary phases limits the realization of the full potential of the existing chiral stationary phases and hampers development of new chiral stationary phases.

Table 2. Classes of Hplc Chiral Stationary Phases

Column chiral selector	Typical mobile phase conditions	Typical analyte features required
pirkle	nonpolar organic; 2-propanol–hexane	π -acid or π -basic moieties for charge transfer complex; hydrogen-bonding or dipole stacking capability near chiral center
protein	phosphate buffers	aromatic near chiral center; organic acids or bases; cationic drugs
cyclodextrin	aqueous buffers; polar organic	good "fit" between chiral cavity or chiral mouth of cyclodextrin and hydro-phobic moiety; hydrogen-bonding capability near chiral center
ligand exchange	aqueous buffers	α -hydroxy or α -amino acids near chiral center; can do nonaromatic
chiral crown ether	0.01 N perchloric acid	primary amines near chiral center; can do nonaromatic
macrocyclic antibiotics	aqueous buffers, nonpolar and polar organic	amines, amides, acids, esters; aromatic; hydrophobic moiety
cellulosic and amylosic	nonpolar organic	aromatic

Ligand-Exchange Phases

Among the earliest reports of chiral separations by liquid chromatography were based on work done by Davankov using ligand exchange (66). These types of columns are available from Phenomenex, J. T. Baker, and Regis Technologies, Inc. As noted previously in the discussion regarding ligand exchange in tlc, chiral separations by ligand exchange in hplc is accomplished using bidentate amino acid ligands, immobilized on a chromatographic substrate, and a divalent metal cation which participates in the formation of a diastereomeric complex with a bidentate chiral analyte and the ligand. Although almost any amino acid can form the basis for the chiral selector, proline and hydroxyproline exhibit the most widespread utility. Also, although other metals can be used, copper(II) is usually the metal of choice and is added to the aqueous buffer mobile phase.

The dependence of chiral recognition on the formation of the diastereomeric complex imposes constraints on the proximity of the metal binding sites, usually either an hydroxy or an amine α to a carboxylic acid, in the analyte. Principal advantages of this technique include the ability to assign configuration in the absence of standards, enantioresolve nonaromatic analytes, use aqueous mobile phases, acquire a stationary phase with the opposite enantioselectivity, and predict the likelihood of successful chiral resolution for a given analyte based on a well-understood chiral recognition mechanism.

Pirkle Phases

The first commercially available chiral column for liquid chromatography was introduced in 1980. This was the first generation of the "Pirkle phases", named after their originator, and was based on *N*-(3,5-dinitrobenzoyl)phenylglycine which was immobilized on a silica support (67). Of all of the commercially available chiral stationary phases for liquid chromatography, the chiral recognition mechanism for the "Pirkle" phases are among the best understood. Chiral recognition on Pirkle phases is thought to depend upon complimentary interactions between the analyte and the selector. These interactions may be π - π , steric, hydrogen-bonding, or dipole–dipole interactions and contribute to the overall stability of the diastereomeric association complexes that form between the individual enantiomers and the chiral selector in the stationary phase. The π - π interactions arise through the association of aromatic systems with complementary electron withdrawing (eg, nitro) and electron donating (eg, alkyl) substituents. The electron-deficient aromatic system is often referred to as π -acidic, the electron-rich system is usually referred to as π -basic. Three unique interactions emanating from the chiral centers of the analyte and their chiral ligand in the stationary phase, seem to be required for successful chiral recognition. A model invoking three unique points of interaction is sometimes referred to as the *3-point interaction model* first proposed by Dalglish (68). To promote analyte–selector interactions, functional groups are often introduced into the analyte through achiral derivatization. For example, amines may be derivatized with 3,5-dinitrobenzoyl chloride to introduce a π -acid aromatic group to promote diastereomeric complexation with a π -basic (R)-*N*-(2-naphthyl)-alanine chiral selector in the stationary phase. Derivatization often has the additional benefit of enhancing solute solubility.

Nonpolar organic mobile phases, such as hexane with ethanol or 2-propanol as typical polar modifiers, are most commonly used with these types of phases. Under these conditions, retention seems to follow normal phase-type behavior (eg, increased mobile phase polarity produces decreased retention). The normal mobile-phase components only weakly interact with the stationary phase and are easily displaced by the chiral analytes thereby promoting enantiospecific interactions. Some of the Pirkle-types of phases have also been used, to a lesser extent, in the reversed phase mode.

Reciprocity, an important concept introduced by Pirkle (69,70), exploited the notion that analytes that were well resolved using a particular chiral selector would likely be good candidates for chiral selectors to enantioresolve analytes similar to the original chiral selector. For instance, the first generation Pirkle phase incorporating *N*-(3,5-dinitrobenzoyl)phenylglycine was very successful at enantioresolving compounds containing naphthyl moieties near the stereogenic center. This insight spawned a second generation of Pirkle phases based on *N*-(2-naphthyl)- α -amino acids (71). These phases were very successful at enantioresolving analytes containing a 3,5-dinitrobenzoyl group, such as 3,5-dinitrophenyl carbamates, and ureas of chiral alcohols and amines (72). These columns are available through a variety of sources including Phenomenex, Regis Technologies, Inc., J. T. Baker, Inc., and Supelco, Inc.

The structure of the Whelk-O-1 phase, the most recent addition to this type of chiral stationary phase, is illustrated in Figure 5. This selector has a wedge-like chiral surface with one edge offering the π -basic tetrahydrophenanthrene ring system; the other edge is comprised of a 3,5-dinitrobenzoyl π -acidic moiety. The amide linkage between the two ring systems presents dipole stacking and hydrogen-bonding interaction sites. The presence of both π -acid and π -base features, as well as the inherent rigidity of the chiral selector, confers greater versatility than any of the previous Pirkle-type phases, imposing fewer constraints on both analyte structural features required for successful enantioresolution and mobile phase conditions. Indeed, this chiral stationary phase has demonstrated considerable chiral selectivity for naproxen, warfarin, and its *p*-chloro analogue under nonaqueous reversed-phase

conditions (73) and reversed-phase conditions (74,75). An additional advantageous feature of this phase is its availability with either the (*R,R*) or (*S,S*) configuration, thus, permitting the enantiomeric elution order to be readily changed. The small size of the chiral selector also promotes fairly high bonded ligand densities in the stationary phase, which coupled with the high enantioselectivities often achieved with these phases, facilitates their use for preparative-scale separations (76).

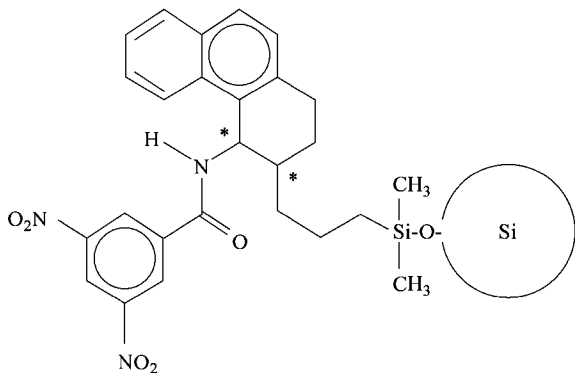


Fig. 5. The structure of the chiral selector in the Whelk-O-1 chiral stationary phase.

Cyclodextrin Phases

Cyclodextrins are macrocyclic compounds comprised of D-glucose bonded through 1,4- α -linkages and produced enzymatically from starch. The greek letter which proceeds the name indicates the number of glucose units incorporated in the CD (eg, α 6, β 7, γ 8, etc). Cyclodextrins are toroidal shaped molecules with a relatively hydrophobic internal cavity (Fig. 6). The exterior is relatively hydrophilic because of the presence of the primary and secondary hydroxyls. The primary C-6 hydroxyls are free to rotate and can partially block the CD cavity from one end. The mouth of the opposite end of the CD cavity is encircled by the C-2 and C-3 secondary hydroxyls. The restricted conformational freedom and orientation of these secondary hydroxyls is thought to be responsible for the chiral recognition inherent in these molecules (77).

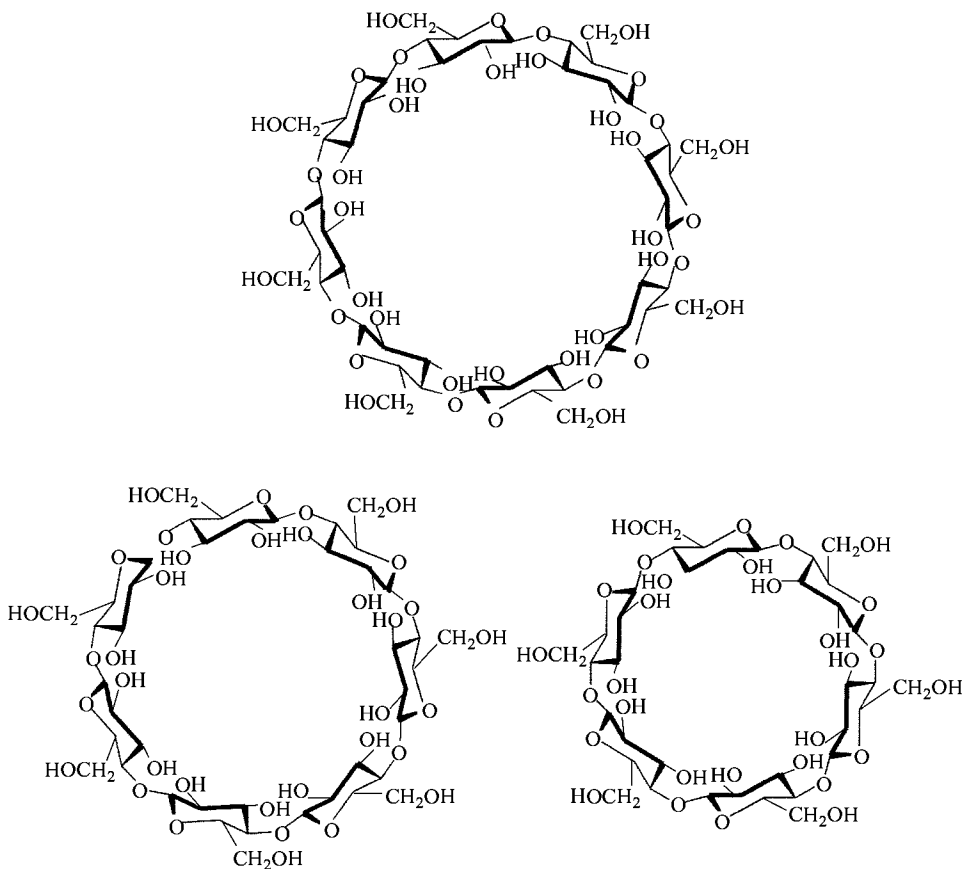


Fig. 6. The structure of the three most common cyclodextrins.

Among the most successful of the liquid chromatographic reversed-phase chiral stationary phases have been the cyclodextrin-based phases, introduced by Armstrong (78,79) and commercially available through Advanced Separation Technologies, Inc. or Alltech Associates. The most commonly used cyclodextrin in hplc is the β -cyclodextrin. In the bonded phases, the cyclodextrins are thought to be tethered to the silica substrate through one or two spacer ligands (Fig. 7). The mechanism thought to be responsible for the chiral selectivity observed with these phases is based on the formation of an inclusion complex between the hydrophobic moiety of the chiral analyte and the hydrophobic interior of the cyclodextrin cavity (Fig. 8). Preferential complexation between one optical isomer and the cyclodextrin through stereospecific interactions with the secondary hydroxyls which line the mouth of the cyclodextrin cavity results in the enantiomeric separation. Unlike the Pirkle-type phases, enantiospecific interactions between the analyte and the cyclodextrin are not the result of a single, well-defined association, but more of a statistical averaging of all the potential interactions with each interaction weighted by its energy or strength of interaction (80).

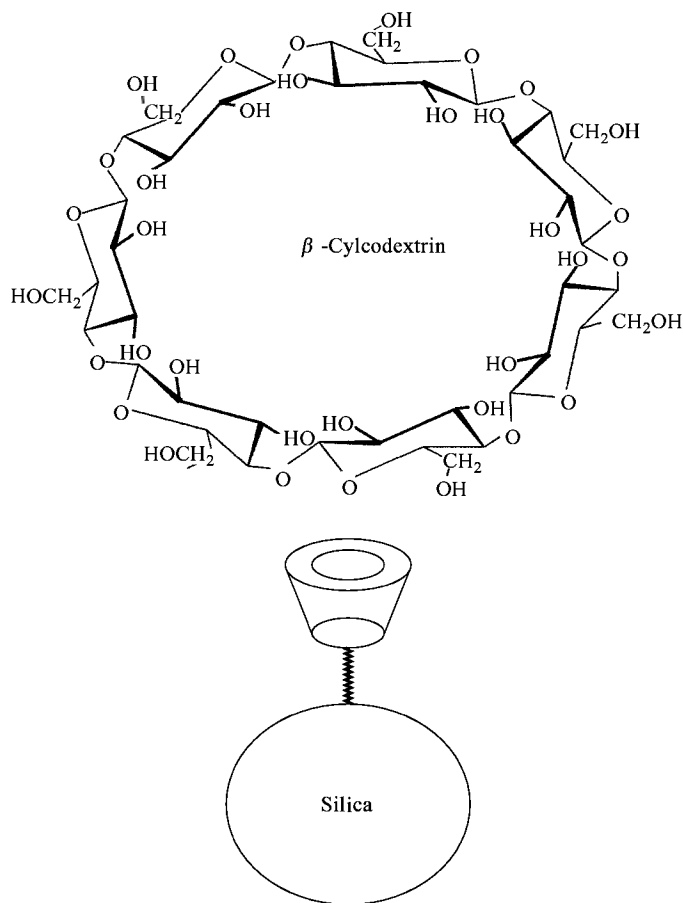


Fig. 7. A tethered cyclodextrin and the structure of β -cyclodextrin, the most common cyclodextrin used as a bonded ligand in liquid chromatography.

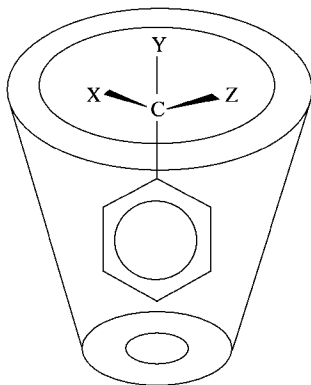


Fig. 8. A hydrophobic inclusion complex between a chiral analyte and a cyclodextrin.

Vast amounts of empirical data suggest that chiral recognition on cyclodextrin phases in the reversed phase mode require the presence of an aromatic moiety that can fit into the cyclodextrin cavity, that there be hydrogen bonding groups in the molecule, and that the hydrophobic and hydrogen-bonding moieties should be in close proximity to the stereogenic center. Chiral recognition seems to be enhanced if the stereogenic center is positioned between two π -systems or incorporated in a ring.

Most of the chiral separations reported to date using the native cyclodextrin-based phases have been accomplished in the reversed-phase mode using aqueous buffers containing small amounts of organic modifiers. However, polar organic mobile phases have gained in popularity recently because of their ease of removal from the sample and reduced tendency to accelerate column degradation relative to the hydroorganic mobile phases (81). In these cases, because the more nonpolar component of the mobile phase is thought to occupy the cyclodextrin cavity, the analyte is thought to sit atop the mouth of the cyclodextrin much like a "lid".

Limitations with the chiral selectivity of the native cyclodextrins fostered the development of various functionalized cyclodextrin-based chiral stationary phases, including acetylated (82,83), sulfated (84), 2-hydroxypropyl (85), 3,5-dimethylphenylcarbamoylated (86) and 1-naphthylethylcarbamoylated (87) cyclodextrin. Each of the glucose residues contribute three hydroxyl groups to which a substituent may be appended; thus, each cyclodextrin contributes multiple sites for derivatization. Typical degrees of substitution per β -cyclodextrin (with 21 hydroxyls) range from three to ten. Hence, there are many residual hydroxyls on each cyclodextrin.

The substituents of these functionalized cyclodextrins seem to play a variety of roles in enhancing chiral recognition. In some cases, the substituent may only serve to enlarge the chiral cavity or may provide alternative interaction sites. For instance, in the case of the naphthylethylcarbamoylated cyclodextrin, the naphthyl ring provides a π -basic site and the carbamate linkage provides additional hydrogen bonding and dipole interaction sites not available with the native cyclodextrin. On the sulfated cyclodextrin phase, the sulfate group presents the potential for ion-pair formation unavailable with the native cyclodextrin. The introduction of 2-hydroxypropyl- and 1-naphthylethylcarbamoyl substituents incorporates additional stereogenic centers onto the cyclodextrin. In some cases, the configuration of the substituent dominated the enantiomeric elution order. However, in other cases, the enantiomeric elution order was independent of the configuration of the substituent. In addition, in some cases, the chiral selectivity of the cyclodextrin seemed to synergistically augment the chiral selectivity of one configuration of the substituent while antagonizing the chiral selectivity of the oppositely configured substituent. A particularly attractive feature of these functionalized cyclodextrins is that many of them exhibit enantioselectivity under hydroorganic reversed phase, as well as normal phase and polar organic mobile-phase conditions, and that each set of conditions can provide chiral separations for analytes which are not resolved under any of the other type of mobile-phase conditions. Further, the chromatographic mode (eg, reversed phase to normal

phase) can be readily changed with no deleterious impact on chiral recognition as long as routine care is taken to avoid problems with solvent immiscibility. The naphthylethyl-carbamoylated cyclodextrin phase was considered to be one of the first "multimodal" chiral stationary phases (88).

Cellulosic and Amylosic Phases

Cellulose and amylose are comprised of the same glucose subunits as the cyclodextrins. In the case of cellulose, the glucose units are attached through 1,4- β -linkages resulting in a linear polymer. In the case of amylose, the 1,4- α -linkages, as are found in the cyclodextrins, are thought to confer helicity to the polymeric chain.

As mentioned previously, cellulosic phases as well as amylosic phases have also been used extensively for enantiomeric separations more recently (89,90). Most of the work in this area has been with various derivatives of the native carbohydrate. The enantioresolving abilities of the derivatized cellulosic and amylosic phases are reported to be very dependent on the types of substituents on the aromatic moieties that are appended onto the native carbohydrate (91). Table 3 lists some of the cellulosic and amylosic derivatives that have been used. These columns are available through Chiral Technologies, Inc. and J. T. Baker, Inc.

Table 3. Carbohydrate Derivatives Used as Hplc Chiral Stationary Phases

Cellulosic	Amylosic
triacetate	
tribenzoate	
tribenzylether	
tricinamate	
triphenylcarbamate	triphenylcarbamate
tris-3,5-dichlorophenylcarbamate	
tris-3,5-dimethylphenylcarbamate	tris-3,5-dimethylphenylcarbamate
tris-1-phenylethylcarbamate	tris-1-phenylethylcarbamate

With the exception of the microcrystalline cellulose I triacetate (92) and tribenzoate materials, which are sufficiently robust to be used directly as packing material, most of the commercially available cellulosic and amylosic phases are comprised of mixtures of exhaustively derivatized polymers which are coated onto large pore γ -aminopropyl silica. These coated polymeric phases exhibit admirable enantioselectivities, but as is the case with all commercially available chiral stationary phases, they also have some potential disadvantages. The large polymer size requires the use of fragile, large pore silica. The fact that the chiral selector for these phases is coated onto the silica sometimes restricts the types of mobile phases that can be used. In addition, the secondary structure of the polymer, which seems to be important in the chiral recognition mechanism, may be altered irreversibly by storing the columns in polar solvents leading to disastrous consequences for chiral separations. The polymeric nature of the chiral selectors for these phases and the importance of the secondary structure also hamper the development of models for the chiral recognition mechanism for these phases (93). Despite these factors, the cellulosic and amylosic phases have enjoyed tremendous success at enantioresolving structurally diverse compounds (94,95) and have some of the highest capacities of all the chiral commercially available chiral stationary phases, thus, rendering them among the most suitable for preparative chromatography (96).

The chiral recognition sites on these polymeric carbohydrate phases are thought to be channels or grooves in the polymer matrix and that analytes are included into these channels. Evidence for inclusion is provided by the enhanced chiral recognition observed for many analytes as the steric bulk of the alcohol mobile phase modifier increases. Chiral recognition seems to require the presence of an aromatic ring, for π - π interactions, and polar sites of unsaturation or hydrogen bonding functionalities. As in the case of the Pirkle-type phases, these chiral stationary phases are usually used in the normal phase mode and mobile phases typically consist of hexane and 2-propanol although there have been some reports of these phases being used in the reversed phase mode (97).

Protein-Based Phases

Proteins, amino acids bonded through peptide linkages to form macromolecular biopolymers, used as chiral stationary phases for hplc include bovine and human serum albumin, α_1 -acid glycoprotein, ovomucoid, avidin, and cellobiohydrolase. The bovine serum albumin column is marketed under the name Resolvosil and can be obtained from Phenomenex. The human serum albumin column can be obtained from Alltech Associates, Advanced Separation Technologies, Inc., and J. T. Baker. The α_1 -acid glycoprotein and cellobiohydrolase can be obtained from Advanced Separation Technologies, Inc. or J. T. Baker, Inc.

In most cases, the protein is immobilized onto γ -aminopropyl silica and covalently attached using a cross-linking reagent such as *N,N'*-carbonyldiimidazole. The tertiary structure or three dimensional organization of proteins are thought to be important for their activity and chiral recognition. Therefore, mobile phase conditions that cause protein "denaturation" or loss of tertiary structure must be avoided.

Typically, the mobile phases used with the protein-based chiral stationary phases consist of aqueous phosphate buffers (98). Often small amounts of organic modifiers, such as methanol, ethanol, propanol, or acetonitrile, are added to reduce hydrophobic interactions with the analyte and to improve enantioselectivity. In some cases, dramatic changes in chiral recognition occur when small amounts of organic modifiers, such as *N,N*-dimethyloctylamine or octanoic acid are added to the mobile phase. It is thought that these additives may be playing an active role in enhancing chiral recognition through absorption of the organic modifier onto the protein which induces conformational changes in the overall tertiary structure of the protein. In these cases, the *allosteric* modifier-mediated conformational changes in the protein are thought to enhance chiral recognition by a variety of mechanisms including changes in the accessibility of various stereospecific sites on the protein or obstruction of nonstereospecific sites (99).

As in the case of the cyclodextrin and amylosic and cellulosic phases, the chiral recognition mechanism for these protein-based phases is not well understood. In some cases, it is thought that analytes may form inclusion complexes with hydrophobic pockets within the biopolymeric matrix. These hydrophobic interactions may couple with hydrogen bonding, electrostatic interactions, and π - π or dipole stacking to individual amino acid residues, thus, contributing to stereospecific orientational constraints within the hydrophobic pockets. Optimization of chromatographic conditions and selection of analytes that can be successfully resolved on these phases is usually done empirically. In addition, the large molecular weight of these biopolymers dictates that the amount of chiral selector that can be immobilized on the column packing material is very small. Although the protein is large, relative to the analytes, the actual region of the protein that affects the chiral separation may be very small. Thus, the capacity (amount of material resolvable during a single chromatographic run) of these columns is generally fairly small ($< \sim 0.1$ mg) and the columns are easily overloaded.

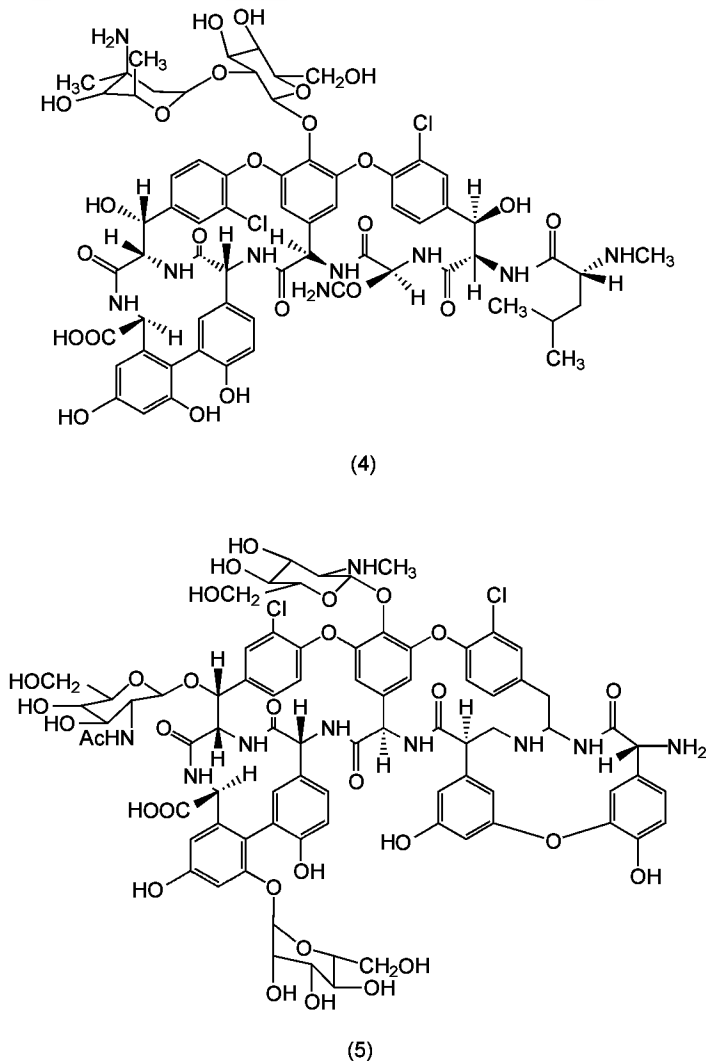
An interesting application of the protein-based phases is various protein binding and displacement experiments which can be done fairly routinely (100). For instance, the chiral selectivity of chiral stationary phases derived from the serum albumin, one of the most abundant blood proteins which functions as a transport protein, from different animal species including rabbit, rat and human has been compared (101). This work suggests that differences in the enantioselectivity, toward a particular drug, of a column derived from human serum albumin and a column derived from some other animal serum albumin might be indicative that a particular species might not be a good animal model during drug development, thus, obviating the need for animal testing.

Chiral separations on protein-based phases may also provide useful information on drug interactions. For instance, the effect of the individual enantiomers of warfarin on the enantioselectivity of human serum albumin toward benzodiazepinones has been studied using a human serum albumin

column with warfarin as a mobile phase additive (102).

Chirobiotic Phases

The chirobiotic chiral stationary phases (103,104) are based on macrocyclic antibiotics such as vancomycin (4) and teicoplanin (5).



These chiral selectors, originally used as chiral additives in capillary zone electrophoresis, incorporate aromatic and carbohydrate, as well as peptide and ionizable moieties. The presence of aromatic groups, allowing for π - π interactions, and the macrocyclic rings, offering potential inclusion complexation, give these phases some of the advantages of the protein-based phases (eg, peptide and hydrogen bonding sites) and the carbohydrate-based phases but with greater sample capacity and greater mobile phase flexibility. Indeed, these phases seem to be truly "multimodal" in that they have demonstrated chiral selectivity in the normal, polar organic, and reversed-phase modes. In the normal and polar organic phase modes, π - π interactions, and dipole stacking are thought to play a predominant role in chiral selector-analyte interactions. In the reversed-phase mode, hydrogen bonding, inclusion complexation and, for charged analytes, electrostatic interactions are thought to dominate the interactions. In addition, the use of such well-defined chiral selectors facilitate method development and optimization. These columns are commercially available through Advanced Separation Technologies, Inc. and Alltech Associates.

Chiral Crown Ether Phases

Chiral crown ethers based on 18-crown-6 (Fig. 9) can form inclusion complexes with ammonium ions and protonated primary amines. Immobilization of these chiral crown ethers on a chromatographic support provides a chiral stationary phase which can resolve most primary amino acids, amines, and amino alcohols. However, the stereogenic center must be in fairly close proximity to the primary amine for successful chiral separation (105,106). Significantly, the chiral crown ether phase is unique in that it is one of the few liquid chromatographic chiral stationary phases that does not require the presence of an aromatic ring to achieve chiral separations. Although chiral recognition seems to be enhanced for analytes containing either bulky substituents or aromatic groups near the stereogenic center, only the presence of the primary amine is mandatory.

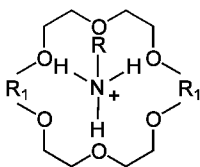


Fig. 9. An inclusion complex formed between a protonated primary amine and a chiral crown ether.

Mobile phases used with this stationary phase are typically 0.01 *N* perchloric acid with small amounts of methanol or acetonitrile. One significant advantage of these phases is that both configurations of the chiral stationary phase are commercially available and can be obtained from J. T. Baker Inc. and Chiral Technologies, Inc. (Crownpak CR).

Chiral Synthetic Polymer Phases

Chiral synthetic polymer phases can be classified into three types. In one type, a polymer matrix is formed in the presence of an optically pure compound to molecularly *imprint* the polymer matrix (Fig. 10) (107,108). Subsequent to the polymerization, the chiral template is removed, leaving the polymer matrix

with chiral cavities. The degree of cross-linking in the polymer matrix and degree of association between the template molecule and the monomer, is governed by the type and concentration of the monomer, the concentration of the template, the solvent and temperature or pressure under which polymerization takes place. All play a role in the chiral selectivities achieved with these phases. The selectivities achieved with these phases are generally excellent, thus, facilitating semipreparative separations. However, the applicability of these chiral stationary phases are generally limited to the analyte upon which the phase is based and a limited number of analogues. In addition, these types of phases generally exhibit poor efficiency in large part because the polymeric matrix contributes to nonsterespecific binding. Advantages of this approach include the ability to prepare reciprocal phases and the predictability of the enantiomeric elution order.

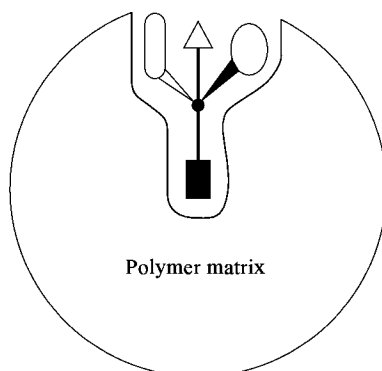


Fig. 10. The relationship between a chiral template molecule and the polymeric matrix formed in the presence of the template molecule.

Another type of synthetic polymer-based chiral stationary phase is formed when chiral catalyst are used to initiate the polymerization. In the case of poly(methyl methacrylate) polymers, introduced by Okamoto, the chirality of the polymer arises from the helicity of the polymer and not from any inherent chirality of the individual monomeric subunits (109). Columns of this type (eg, Chiralpak OT) are available from Chiral Technologies, Inc., or J. T. Baker Inc.

A third type of synthetic polymer-based chiral stationary phase, developed by Blaschke (110), is produced when a chiral selector is either incorporated within the polymer network (111) or attached as pendant groups onto the polymer matrix. Both are analogous to methods used to produce polymeric chiral stationary phases for gc. The polymers can be either coated onto a silica substrate, comonomers bearing silane functional groups may be added for subsequent reaction with the silica, or the silica may be chemically modified to incorporate monomer-bearing silanes. More recently, L-valine-3,5-dimethylanilide has been bonded to a poly(glycidylmethacrylate-*co*-ethylenedimethacrylate) polymer which formed the underlying substrate (112). Chemical bonding of the polymer to the substrate eases the mobile phase restrictions imposed on the coated chiral polymer stationary phases.

In general, the synthetic polymeric phases seem to have polarities analogous to diol-type phases and a wide range of mobile phase conditions have been used including hexane, various alcohols, acetonitrile, tetrahydrofuran, dichloromethane and their mixtures, as well as aqueous buffers.

Chiral Separation Validation for Hplc

Chiral separations present special problems for validation. Typically, in the absence of spectroscopic confirmation (eg, mass spectral or infrared data), conventional separations are validated by analyzing "pure" samples under identical chromatographic conditions. Often, two or more chromatographic stationary phases, which are known to interact with the analyte through different retention mechanisms, are used. If the pure sample and the unknown have identical retention times under each set of conditions, the identity of the unknown is assumed to be the same as the pure sample. However, often the chiral separation that is obtained with one type of column may not be achievable with any other type of chiral stationary phase. In addition, "pure" enantiomers are generally not available.

Most commonly, uv or uv-vis spectroscopy is used as the basis for detection in hplc. When using a chiral stationary phase, confirmation of a chiral separation may be obtained by either monitoring the column effluent at more than one wavelength or by running the sample more than once. The same mobile-phase conditions are used, but monitoring is done at different wavelengths. Because enantiomers have identical spectra in an achiral environment, the ratio of the peaks for the two enantiomers should be independent of wavelength. Although not absolute proof of a chiral separation, this approach does provide strong supporting evidence.

As in tlc, another method to validate a chiral separation is to collect the individual peaks and subject them to some type of optical spectroscopy, such as, circular dichroism or optical rotary dispersion. Enantiomers have mirror image spectra (eg, the negative maxima for one enantiomer corresponds to the positive maxima for the other enantiomer). One problem with this approach is that the analytes are diluted in the mobile phase. Thus, the sample must be injected several times. The individual peaks must be collected and subsequently concentrated to obtain adequate concentrations for spectral analysis.

Alternatively, a chiroptical spectroscopy can be used as the basis for detection on-line using commercially available optical rotary dispersion or circular dichroism-based detectors. Optical rotary dispersion instruments are analogous to refractive index-based detectors for conventional chromatography in that they are universal, do not require the presence of a chromophore in the analyte and have the least sensitivity of the optical detectors. Circular dichroic detection, although more sensitive than optical rotary dispersion-based detection, requires not only the presence of a uv chromophore in the analyte, but that the chromophore be not too distant from the asymmetric center of the analyte. Figure 11a illustrates a simulated chromatogram for an enantiomeric separation obtained using a conventional absorption detector. Figure 11b illustrates a simulated chromatogram for an enantiomeric separation using circular dichroic detection. Both types of chiroptical detectors produce positive and negative peaks for the two enantiomers. However, neither chiroptical detector can distinguish a fair separation (Fig. 11d) from a poor separation (Fig. 11f) in which there is considerable overlap of the two peaks. This is because the signals generated by the two enantiomers have opposite signs, and thus, any overlap causes cancellation of signal. Further, peak overlap results in nonlinear detector response vs concentration. Therefore, some other detection method must be used in conjunction with either of these types of detection. Nevertheless, as can be seen from Figure 11f, chiroptical detection can be advantageous if there is considerable overlap of the two peaks. In this case, chiroptical detection may reveal that the leading and trailing edges of the peak are enantiomerically enriched which may not be apparent from the chromatogram obtained with nonchiroptical detection (Fig. 11e).

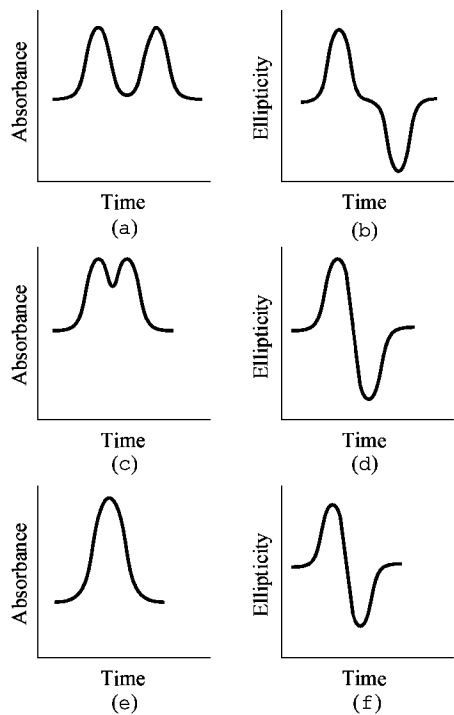


Fig. 11. Simulated chromatograms of chiral separations obtained using nonchiroptical detection (a, c, e) and chiroptical detection (b, d, f) illustrating the effect of peak overlap on the resultant chromatogram.

Another method for validating chiral separations by lc is to couple the chromatographic system to a mass spectrometer. In mass spectrometry, high energy ions are used to bombard molecules exiting the column. The impact of the high energy ions causes the molecules to "fragment" into various ions which are then sent to a "mass discriminator". The ion fragments are detected and a fragmentation pattern or mass spectrum is reconstructed. Enantiomers have identical fragmentation patterns. Hence, identical fragmentation patterns for two peaks in the chromatogram confirms a chiral separation.

Chiral Stationary Phases for Gas Chromatography

Although chiral stationary phases for gas chromatography (gc) were introduced before liquid chromatographic chiral stationary phases, development of gc chiral stationary phases lagged behind for a variety of reasons. First of all, analysis by gc requires that the analyte be volatile and thermally stable. This condition often requires that the analyte be derivatized with an achiral reagent prior to chromatographic analysis to enhance sample volatility. In some cases, derivatization may actually enhance detector response (eg, trifluoroacetylation amplifies electron capture detection) or chiral interactions. However, it should be noted that the presence of more than one type of functionality (eg, amine and alcohol) in the analyte with differing reactivities toward the derivatizing agent may add additional complications. Typical achiral derivatizing reagents, as well as the appropriate functionality required in the analyte, are listed in Table 4.

Table 4. Analyte Functional Groups and Typical Achiral Derivatizing Reagents

Type of derivatizing agent	Analyte functional group	Examples of derivatizing agents
alkyl silyl	alcohols	<i>N</i> -trimethylsilylimidazole
	thiols	
	carboxylic acid	
	amines	
acyl, haloacyl or anhydride	alcohols	<i>N,O</i> -bis(trimethylsilyl)-trifluoroacetamide acetic acid heptafluorobutyl chloride trifluoroacetyl chloride
	amines	
	amides	
	oximes	
	thiols	
	ketones	
	carboxylic acids	
alcohol	carboxylic acids	methanol
alkyl halides	carboxylic acids	methyl bromide
dialkyl	carboxylic acids	diazomethane
isocyanate	sulfonic acids	isopropylisocyanate
	phenols	
	alcohols	
	amines	
phosgene	hydroxy acids	
	β-amino alcohols	
	β-amino thiols	
	diols	
	<i>N</i> -methylamino acids	
alkyl hydroxylamine	ketones	methylhydroxylamine

The use of gas chromatography for chiral separations was also hampered because the high column temperatures typically used in gc tend to accelerate racemization of the stationary phase, thus, decreasing column longevity. The high column temperatures typically used in gc also tend to accelerate racemization of the analyte. In addition, the differences in the stabilities of the diastereomeric complexes formed between the enantiomers and the stationary phase tends to be overcome by the high column temperatures. Finally, preparative scale separations are generally harder to implement in gc than in hplc. However, the inability of most liquid chromatographic methods to resolve chirally small nonaromatic compounds that are frequently used as chiral synthetic building blocks, as well as improvements in gc column technology, has led to renewed interest in chiral stationary phases for gc.

Gc chiral stationary phases can be broadly classified into three categories: diamide, cyclodextrin, and metal complex.

Diamide Chiral Separations. The first chiral stationary phase for gas chromatography was reported by Gil-Av and co-workers in 1966 (113) and was based on *N*-trifluoroacetyl (*N*-TFA) *L*-isoleucine lauryl ester coated on an inert packing material. It was used to resolve the trifluoroacetylated derivatives of amino acids. Related chiral selectors used by other workers included *n*-dodecanoyl-*L*-valine-*t*-butylamide and *n*-dodecanoyl-(*S*)- α -(1-naphthyl)-ethylamide. The presence of the long alkyl groups lowered chiral selector volatility thus reducing but not entirely eliminating column bleed and improving column longevity.

The first commercially available chiral column was the Chiralsil-val (Fig. 12), which was introduced in 1976 (114) for the separation of amino acid-type compounds by gas chromatography. It is based on a polysiloxane polymer containing chiral side chains incorporating *L*-valine-*t*-butylamide. The polysiloxane backbone improved the thermal stability of these chiral stationary phases relative to the original coated columns and extended the operating temperatures up to 220°C. The column is effective for the separation of perfluoroacetylated and esterified amino acids, amino alcohols, and some chiral sulfoxides. Another polysiloxane-based chiral stationary phase incorporating *L*-valine-(*R*)- α -phenylethylamide appended onto hydrolyzed XE-60 was found to be particularly successful at resolving perfluoroacetylated amino alcohol derivatives (115). Through judicious choice of derivatizing agent, chiral separations were obtained for a wider range of compounds, including amino alcohols, α -hydroxy acids, diols and ketones, than had previously been obtainable using these types of stationary phases (116).

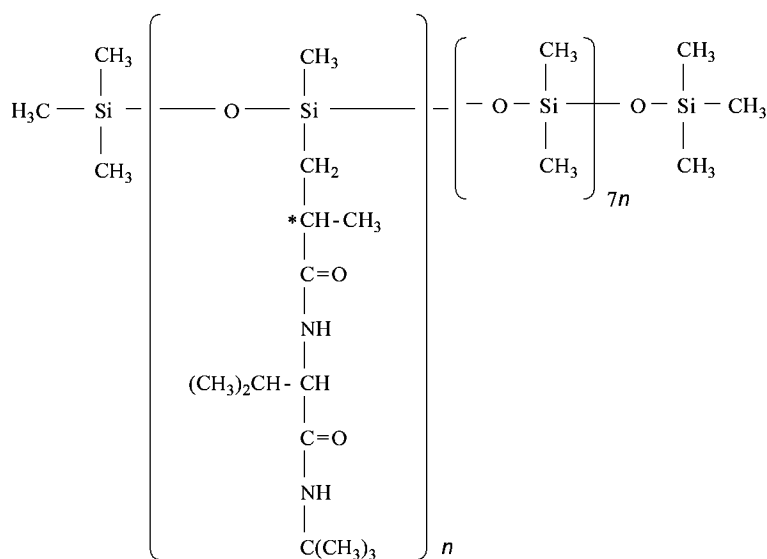
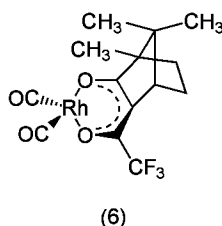


Fig. 12. Structure of Chiralsil-val.

The chiral recognition mechanism for these types of phases was attributed primarily to hydrogen bonding and dipole–dipole interactions between the analyte and the chiral selector in the stationary phase. It was postulated that chiral recognition involved the formation of transient five- and seven-membered association complexes between the analyte and the chiral selector (117).

On each of these amino acid-based chiral stationary phases, the configuration of the most retained enantiomer corresponds to the configuration of the chiral selector in the stationary phase (eg, *L*-amino acids were retained longer on the *N*-trifluoroacetyl (*N*-TFA) *L*-isoleucine lauryl ester column). Thus, configuration of the analytes can be assigned even if no optically pure material is available as long as optically pure standard materials are available for structurally related compounds. Another advantage is that stationary phases incorporating the chiral selector with either configuration may be readily prepared or are commercially available. Thus, the elution order may readily be reversed by using a column containing the chiral selector with the opposite configuration, thus, providing another tool for chiral separation validation. Also, quantitation of a very small peak in the presence of a very large peak is generally easier if the smaller peak elutes first.

Metal Complex. Complexation gas chromatography was first introduced by V. Schurig in 1980 (118) and employs transition metals (eg, nickel, cobalt, manganese or rhodium) complexed with chiral terpenoid ketoenolate ligands such as 3-trifluoroacetyl-1*R*-camphorate (6), 1*R*-3-pentafluoro-benzoylcamphorate or 3-heptafluorobutanoyl-(1*R*,2*S*)-pinnanone-4-ate. In most cases, the chiral selector is dissolved in a polymer matrix coated on the interior walls of a capillary column. This class of chiral columns is particularly adept at enantioresolving some olefins and oxygen-containing compounds such as ketones, ethers, alcohols, spiroacetals, oxiranes, and esters. Many of these compounds lack suitable functionality for derivatization with chiral reagents, and thus, are not amenable to diastereomer formation. Unfortunately, as is the case with many of the chiral stationary phases, the chiral recognition mechanism is not sufficiently refined to allow for prediction of analyte absolute configuration on the basis of retention times except for a very limited number of cases. Nevertheless, these columns allow for the direct chiral separation of compounds that are important synthetic precursors and may be difficult to separate by any other method.



Cyclodextrins. As indicated previously, the native cyclodextrins, which are thermally stable, have been used extensively in liquid chromatographic chiral separations, but their utility in gc applications was hampered because their high crystallinity and insolubility in most organic solvents made them difficult to formulate into a gc stationary phase. However, some functionalized cyclodextrins form viscous oils suitable for gc stationary-phase coatings and have been used either neat or diluted in a polysiloxane polymer as chiral stationary phases for gc (119). Some of the derivatized cyclodextrins which have been adapted to gc phases are 3-*O*-acetyl-2,6-di-*O*-pentyl, 3-*O*-butyryl-2,6-di-*O*-pentyl, 2,6-di-*O*-methyl-3-*O*-trifluoroacetyl, 2,6-dipentyl, 2-*O*-methyl-3,6-di-*O*-pentyl, permethyl, permethylhydroxypropyl, perpentyl, and propionyl. Several of these are available commercially. For instance, Advanced Separation Technologies, Inc., Alltech Associates, J. & W. Scientific, and Supelco, Inc., all carry cyclodextrin-based chiral gc columns. Although these derivatized cyclodextrins are often coated, neat, onto the capillary column inner walls, some work has been done to tether the cyclodextrins to a polysiloxane backbone to enhance the thermal stability of the resultant phase (120). Some of the separations obtained with these materials are quite remarkable and include compounds such as halogenated alkanes (Fig. 13) (121), alcohols, alkenes, bicyclic compounds, and simple alkanes.

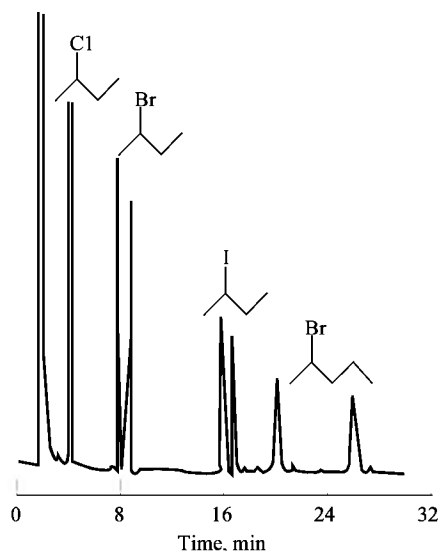


Fig. 13. Enantiomeric separations of monohalohydrocarbons on a 2,6-*O*-dipentyl-3-*O*-trifluoroacetyl- γ -cyclodextrin coated capillary column (10 m, 0.25 mm ID). Column temperature, 30°C; nitrogen carrier gas, 20.7 kPa (3 psi).

Although the chiral recognition mechanism of these cyclodextrin-based phases is not entirely understood, thermodynamic and column capacity studies indicate that the analytes may interact with the functionalized cyclodextrins by either associating with the outside or mouth of the cyclodextrin, or by forming a more traditional inclusion complex with the cyclodextrin (122). As in the case of the metal-complex chiral stationary phase, configuration assignment is generally not possible in the absence of pure chiral standards.

Chiral Separation Validation for Gas Chromatography

The special problems for validation presented by chiral separations can be even more burdensome for gc because most methods of detection (eg, flame ionization detection or electron capture detection) in gc destroy the sample. Even when nondestructive detection (eg, thermal conductivity) is used, individual peak collection is generally more difficult than in lc or tlc. Thus, off-line chiroptical analysis is not usually an option. Fortunately, gc can be readily coupled to a mass spectrometer and is routinely used to validate a chiral separation.

Conclusions

The field of chiral separations has grown explosively into a well-developed specialty within separation science in the last two decades. Considerable effort in the field has thus far been directed toward solving analytical separation problems and in the design and development of new chiral separation methods. However, in the future, chiral separations may have the potential to provide an invaluable tool for probing intermolecular interactions, increasing our understanding of the chemistry behind some biological processes and disease states (123,124). Also the efficacy of drug therapy can be enhanced by eliminating or minimizing untoward side effects (125) produced by distomeric ballast and allowing for better control in the application of chiral or prochiral entities to biological or environmental systems.

BIBLIOGRAPHY

1. L. Pasteur, *Comptes Rendus de l'Academie des Sciences* **26**, 535 (1848).
2. W. J. Wechter, D. G. Loughhead, R. J. Reischer, G. J. Van Giessen, and D. G. Kaiser, *Biochem. Biophys. Res. Comm.* **61**, 833 (1974).
3. R. A. Kuzel, S. K. Bhasin, H. G. Oldham, L. A. Damani, J. Murphy, P. Camilleri, and A. J. Hutt, *Chirality* **6**, 607 (1994).
4. S. A. Tobert, *Clin. Pharmacol. Ther.* **29**, 344 (1981).
5. H. Ohkawa, *Bull. Environ. Contamin. Toxicol.* **18**, 534 (1977).
6. H. G. Schmarr, A. Mosandl, and K. Grob, *Chromatographia* **29**, 125 (1990).
7. T. Aleibi-Kolbah, G. Filix, and I. W. Wainer, *Chromatographia* **35**, 264 (1993).
8. D. W. Armstrong, J. R. Faulkner, Jr., and S. M. Han, *J. Chromatogr.* **452**, 323 (1988).
9. T. J. Ward, *Anal. Chem.* **66**, 632A (1994).
10. C. Pettersson, A. Karlsson, and C. Gioeli, *J. Chromatogr.* **407**, 217 (1987).
11. J. D. Duncan, D. W. Armstrong, and A. M. Stalcup, *J. Liq. Chromatogr.* **13**, 1091 (1990).
12. D. W. Armstrong, F.-Y. He, and S. M. Han, *J. Chromatogr.* **448**, 345 (1988).
13. A. D. Cooper, T. M. Jeffries, and R. M. Gaskell, *Anal. Proc.* **29**, 258 (1992).
14. D. W. Armstrong and Y. Zhou, *J. Liq. Chromatogr.* **17**, 1695 (1994).
15. C. Pettersson and G. Schill, *J. Liq. Chromatogr.* **9**, 269 (1986).
16. M.-B. Huang, H.-K. Li, G.-L. Li, C.-T. Yan, and L.-P. Wang, *J. Chromatogr. A* **742**, 289 (1996).
17. D. W. Armstrong, J. R. Faulkner, Jr., and S. M. Han, *J. Chromatogr.* **452**, 323 (1988).
18. D. W. Armstrong, F.-Y. He, and S. M. Han, *J. Chromatogr.* **448**, 345 (1988).
19. J. D. Duncan, D. W. Armstrong, and A. M. Stalcup, *J. Liq. Chromatogr.* **13**, 1091 (1990).
20. W. G. Kuhr, *Anal. Chem.* **62**, 403R (1990).
21. B. L. Karger, *Amer. Lab.*, 23 (Oct. 1993).
22. J. W. Jorgenson, and K. D. Lukacs, *Anal. Chem.* **53**, 1298 (1981).
23. A. S. Cohen, A. Paulus, and B. L. Karger, *Chromatographia* **24**, 15 (1987).
24. P. Sun, N. Wu, G. Barker, and R. A. Hartwick, *J. Chromatogr.* **648**, 475 (1993).
25. S. Terabe, K. Otsuka, and H. Nishi, *J. Chromatogr.* **666**, 295 (1994).
26. Y. Y. Rawjee and G. Vigh, *Anal. Chem.* **66**, 619 (1994).
27. A. M. Stalcup and K. H. Gahm, *Anal. Chem.* **68**, 1360 (1996).
28. A. M. Stalcup and N. M. Agyei, *Anal. Chem.* **66**, 3054 (1994).
29. H. Soini, M. Stefansson, M.-L. Riekkola, and M. V. Novotny, *Anal. Chem.* **66**, 3477 (1994).
30. P. Gozel, E. Gassman, H. Michelson, and R. N. Zare, *Anal. Chem.* **59**, 44 (1987).
31. A. M. Stalcup and K. H. Gahm, *J. Microcolumn Sep.* **8**, 145 (1996).
32. D. W. Armstrong, K. Rundlett, and G. L. Reid, *Anal. Chem.* **66**, 1690 (1994).

33. T. O. Cole, M. J. Sepaniak, and W. L. Hinze, *J. High Resolut. Chromatogr. Chromatogr. Commun.* **13**, 570 (1990).
34. Y. Mechref and Z. El Rassi, *Chirality* **8**, 515 (1996).
35. N. M. Agyei, K. H. Gahm, and A. M. Stalcup, *Anal. Chim. Acta* **307**, 185 (1995).
36. A. M. Stalcup and K. H. Gahm, *J. Microcol. Sep.* **8**, 145 (1996).
37. R. S. Sahota and M. G. Khaledi, *Anal. Chem.* **66**, 1141 (1994).
38. F. Wang and M. G. Khaledi, *Anal. Chem.* **68**, 3460 (1996).
39. S. A. C. Wren and R. C. Rowe, *J. Chromatogr.* **603**, 235 (1992).
40. A. Guttman, A. Paulus, A. S. Cohen, N. Grinberg, and B. L. Karger, *J. Chromatogr.* **448**, 41 (1988).
41. S. Junsson, A. Schim, R. Isaksson, C. Pettersson, and G. Pettersson, *Chirality* **5**, 505 (1992).
42. A. M. Stalcup, S. Gratz, and Y. Jin, unpublished results.
43. L. Witherow, T. D. Spurway, R. J. Ruane, I. D. Wilson, and K. Longdon, *J. Chromatogr.* **553**, 497–501 (1991).
44. H. T. K. Xuan and M. Lederer, *J. Chromatogr.* **635**, 346 (1993).
45. H. T. K. Xuan and M. Lederer, *J. Chromatogr.* **645**, 185 (1993).
46. C. A. Brunner and I. W. Wainer, *J. Chromatogr.* **472**, 277 (1989).
47. R. Bhushan and V. Parshad, *J. Chromatogr.* **721**, 369 (1996).
48. R. Bhushan and I. Ali, *Chromatographia* **35**, 679 (1993).
49. R. Bhushan and V. Parshad, *J. Chromatogr.* **736**, 235 (1996).
50. U. A. Th. Brinkman and D. Kamminga, *J. Chromatogr.* **330**, 375 (1985).
51. V. Mathur, N. Kanoongo, R. Mathur, C. K. Narang, and N. K. Mathur, *J. Chromatogr.* **685**, 360 (1994).
52. M. Remelli, R. Piazza, and F. Pulidori, *Chromatographia* **32**, 278 (1991).
53. D. W. Armstrong and W. DeMond, *J. Chromatogr. Sci.* **22**, 411 (1984).
54. D. W. Armstrong, Y. Tang, S. Chen, Y. Zhou, C. Bagwill, and J. R. Chen, *Anal. Chem.* **66**, 1473 (1994).
55. W. H. Pirkle, J. M. Finn, B. C. Hamper, J. L. Schreiner, and J. R. Pribish, *Am. Chem. Soc. Symp. Ser. No. 185*, Chapt. 18, 1982.
56. W. H. Pirkle and P. G. Murray, *J. Liq. Chromatogr.* **13**, 2123 (1990).
57. J. Hermansson and M. Eriksson, *J. Liq. Chromatogr.* **9**, 621 (1986).
58. G. Schill, I. W. Wainer, and S. A. Barkin, *J. Liq. Chromatogr.* **9**, 641 (1986).
59. S. Allenmark, *J. Liq. Chromatogr.* **9**, 425 (1986).
60. M. Okamoto and H. Nakazawa, *J. Chromatogr.* **508**, 217 (1990).
61. T. Miwa, H. Kuoda, S. Sakashita, N. Asakawa, and Y. Miyake, *J. Chromatogr.* **511**, 89 (1990).
62. R. Isaksson, P. Erlandsson, L. Hansson, A. Holmberg, and S. Berner, *J. Chromatogr.* **498**, 257 (1990).
63. Y. Okamoto, R. Aburatani, K. Hatano, and K. Hatada, *J. Liq. Chromatogr.* **11**, 2147 (1988).
64. T. Shinbo, T. Yamaguchi, K. Nishimura, and M. Sugiura, *J. Chromatogr.* **405**, 145 (1987).
65. M. Hilton and D. W. Armstrong, *J. Liq. Chromatogr.* **14**, 9 (1991).
66. V. A. Davankov, A. A. Kurganov, and A. S. Bochoy, *Adv. Chromatogr.* **22**, 71 (1983).
67. W. H. Pirkle, J. M. Finn, J. L. Schreiner, and B. C. Hamper, *J. Am. Chem. Soc.* **103**, 3964 (1981).
68. C. E. Dalglish, *J. Chem. Soc.*, 3940 (1952).
69. W. H. Pirkle, D. W. House, and J. M. Finn, *J. Chromatogr.* **192**, 143 (1980).
70. W. H. Pirkle and T. C. Pochapsky, *J. Am. Chem. Soc.* **108**, 352 (1986).
71. W. H. Pirkle, T. C. Pochapsky, G. S. Mahler, D. E. Corey, D. S. Reno, and D. M. Alessi, *J. Org. Chem.* **51**, 4991 (1986).
72. W. H. Pirkle, G. Mahler, and M. H. Hyun, *J. Liq. Chromatogr.* **9**, 443 (1986).
73. W. H. Pirkle and C. J. Welch, *Tetrahedron Asym.* **5**, 777 (1994).
74. C. J. Welch, T. Szczerba, and S. R. Perrin, *J. Chromatogr. A* **758**, 93 (1997).
75. Regis Technologies Application Guide, pp. 30, 31.
76. C. J. Welch and S. R. Perrin, *J. Chromatogr.* **690**, 218 (1995).
77. T. J. Ward and D. W. Armstrong, "Cyclodextrin Stationary Phases" in M. Zief and L. J. Crane, eds., *Chromatographic Chiral Separations*, Marcel Dekker, Inc., New York, 1988, pp. 131.
78. S. M. Han and D. W. Armstrong, in A. M. Krstulovic, ed., *Chiral Separations by HPLC*, John Wiley & Sons, Inc., New York, 1989, pp. 208–287.
79. D. W. Armstrong, T. J. Ward, R. D. Armstrong, and T. E. Beesley, *Science* **232**, 1132 (1986).
80. R. E. Boehm, D. E. Martire, and D. W. Armstrong, *Anal. Chem.* **60**, 522 (1988).
81. D. W. Armstrong, S. Chen, C. Chang, and S. Chang, *J. Liq. Chromatogr.* **15**, 545 (1992).
82. A. M. Stalcup, J. R. Faulkner, Y. Tang, D. W. Armstrong, L. W. Levy, E. Regalado, *Biomed. Chromatogr.* **5**, 3 (1991).
83. P. Camilleri, C. A. Reid, and D. T. Manallack, *Chromatographia* **38**, 771 (1994).
84. A. M. Stalcup and K. H. Gahm, *Anal. Chem.* **68**, 1369 (1996).
85. A. M. Stalcup, S. Chang, D. W. Armstrong, and J. Pitha, *J. Chromatogr.* **513**, 181 (1990).
86. D. W. Armstrong, A. M. Stalcup, M. L. Hilton, J. D. Duncan, J. R. Faulkner, and S. C. Chang, *Anal. Chem.* **62**, 1610 (1990).
87. A. M. Stalcup, S. C. Chang, and D. W. Armstrong, *J. Chromatogr.* **540**, 113 (1991).
88. D. W. Armstrong, M. Hilton, and L. Coffin, *LC-GC* **9**, 647 (1992).
89. H. Hopf, W. Grahm, D. G. Barrett, A. Gerdes, J. Hilmer, J. Hucker, Y. Okamoto, and Y. Kaida, *Chem. Ber.* **123**, 841 (1990).
90. Y. Okamoto, Y. Kaida, R. Aburatani, and K. Hatada, *J. Chromatogr.* **477**, 367 (1989).
91. Y. Okamoto, K. Hatano, R. Aburatani, and K. Hatada, *Chem. Lett.*, 715 (1989).
92. J. M. Jansen, S. Coppinga, G. Gruppen, R. Isaksson, D. T. Witte, and C. J. Grol, *Chirality* **6**, 596 (1994).
93. T. Shibata, I. Okamoto, and K. Ishii, *J. Liq. Chromatogr.* **9**, 313 (1986).
94. D. T. Witte, F. J. Bruggeman, J. P. Franke, S. Coppinga, J. M. Jansen, and R. A. De Zeeuw, *Chirality* **5**, 545 (1993).
95. Y. Okamoto, T. Ohashi, Y. Kaida, and E. Yashima, *Chirality* **5**, 616 (1993).
96. J. Wagner, H.-J. Hamann, W. Düpke, A. Kunath, and E. Huf, *Chirality* **7**, 243 (1995).
97. M. Tanaka, H. Yamazaki, and H. Hokusui, *Chirality* **7**, 612 (1995).
98. J. Iredale, A.-F. Aubry, and I. W. Wainer, *Chromatographia* **31**, 329 (1991).
99. T. A. G. Noctor, I. W. Wainer, and D. S. Hage, *J. Chromatogr.* **577**, 305 (1995).
100. D. S. Hage, T. A. G. Noctor, and I. W. Wainer, *J. Chromatogr. A* **693**, 23 (1995).
101. G. Massolini, A.-F. Aubry, A. McGann, and I. Wainer, *Biochem. Pharm.* **46**, 1285 (1993).
102. E. Domenici, C. Bertucci, P. Salvadori, and I. W. Wainer, *J. Pharm. Sci.* **80**, 164 (1991).
103. D. W. Armstrong, Y. Tang, and S. Chen, *Anal. Chem.* **66**, 473 (1994).
104. D. W. Armstrong, Y. Liu, and K. H. Ekborgott, *Chirality* **7**, 474 (1995).
105. T. Shinbo, T. Yamaguchi, K. Nishimura, and M. Sugiura, *J. Chromatogr.* **405**, 145 (1987).
106. M. Hilton and D. W. Armstrong, *J. Liq. Chromatogr.* **14**, 9 (1991).
107. B. Sellergren, M. Lepistö, and K. Mosbach, *J. Am. Chem. Soc.* **110**, 5853 (1988).
108. L. Fischer, R. Müller, and B. Ekberg, *J. Am. Chem. Soc.* **113**, 9358 (1991).
109. Y. Okamoto, K. Suzuki, K. Ohta, K. Hatada, and H. Yuki, *J. Am. Chem. Soc.* **101**, 4763 (1979).

110. G. Blaschke, W. Bruker, and W. Fraenkel, *Angew. Chem.* **98**, 808 (1986).
111. S. G. Allenmark, S. Andersson, P. Müller, and D. Sanchez, *Chirality* **7**, 248 (1995).
112. Y. Liu, *Anal. Chem.* **69**, 61 (1997).
113. E. Gil-Av, B. Feibush, and R. Charles-Sigler, *Tetrahedron Lett.* **10**, 1009 (1966).
114. H. Frank, G. J. Nicholson, and E. Bayer, *J. Chromatogr. Sci.* **15**, 174 (1974).
115. W. A. König, I. Benecke, and S. Sievers, *J. Chromatogr.* **217**, 71 (1981).
116. W. A. König and E. Steinbach, and K Ernst, *J. Chromatogr.* **301**, 129 (1984).
117. B. Feibush, A. Balan, B. Altman, and Gil-Av, *J. Chem. Soc. Perkin II*, 1230 (1979).
118. V. Schurig, *Chromatographia* **13**, 263 (1980).
119. H. P. Nowotny, D. Schmalzing, D. Wistuba and V. Schurig, *J. High Resolut. Chromatogr. Chromatogr. Commun.* **12**, 383 (1989).
120. D. W. Armstrong, Y. Tang, T. Ward, and M. Nichols, *Anal. Chem.* **65**, 1114 (1993).
121. W.-Y. Li, H. L. Jin, and D. W. Armstrong, *J. Chromatogr.* **509**, 303 (1990).
122. A. Berthod, W. Li, and D. W. Armstrong, *Anal. Chem.* **64**, 873 (1992).
123. D. W. Armstrong, J. Zukowski, N. Ercal, and M. Gasper, *J. Pharm. Biomed. Anal.* **11**, 881 (1993).
124. J. Roboz, E. Nieves, and J. F. Roland, *J. Chromatogr.* **500**, 413 (1990).
125. I. W. Wainer, J. Ducharme, C. P. Granvil, H. Parenteau, and S. Abdullah, *J. Chromatogr. A* **694**, 169 (1995).

General References

C. F. Poole and S. K. Poole, *Chromatography Today*, Elsevier Science Publishers B.V., Amsterdam, The Netherlands, 1991.
I. W. Wainer, *Drug Stereochemistry: Analytical Methods and Pharmacology*, Marcel Dekker, Inc., New York, 1993.
W. A. König, *The Practice of Enantiomer Separation by Capillary Gas Chromatography*, Huthig Verlag, Heidelberg, Germany, 1987.
P. Schreier, A. Bernreuther, and M. Huffer, *Analysis of Chiral Organic Molecules*, Walter de Gruyter & Co., Berlin, Germany, 1995.
G. Subramanian, *A Practical Approach to Chiral Separations by Liquid Chromatography*, VCH, Weinheim, Germany, 1994.

Apryll M. Stalcup
University of Cincinnati

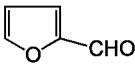
FURAN DERIVATIVES

Furan (1) is a 5-membered heterocyclic, oxygen-containing, unsaturated ring compound. From a chemical perspective it is the basic ring structure found in a whole class of industrially significant products. The furan nucleus is also found in a large number of biologically active materials. Compounds containing the furan ring (as well as the tetrahydrofuran ring) are usually referred to as furans (1). From a manufacturing standpoint, however, furfural (2) is the feedstock from which all of the commercial furan derivatives are derived, as it the easiest and least expensive to manufacture. In this article, the more common name furfural is used in place of the *Chemical Abstracts* name, 2-furancarboxaldehyde.

Furan is produced from furfural commercially by decarbonylation; loss of carbon monoxide from furfural gives furan directly. Tetrahydrofuran (3) is the saturated analogue containing no double bonds.



furan [110-00-9]
(1)



furfural [98-01-1]
(2)



tetrahydrofuran [109-99-9]
(3)

Furfural is derived from biomass by a process in which the hemicellulose fraction is broken down into monomeric 5-carbon sugar units which then are dehydrated to form furfural.

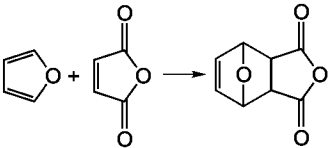
This article is limited primarily to simple furans in which the nucleus occurs as a free monocycle, and in general, does not include compounds in which the furan ring is fused to another ring. In recent years it has mistakenly become common practice to refer to polychlorinated isobenzofurans simply as *furans*. These isobenzofurans are comprised of chlorine-containing fused rings, are not simple furans, and are not included here. The presentation is generally limited to compounds (including resins or polymers) derived from furfural. Ring closure methods to prepare furan or tetrahydrofuran compounds are not covered.

Tetrahydrofuran (3) is produced commercially from furfural by decarbonylation followed by hydrogenation; it is also produced by several different methods from other raw materials. A complete discussion of tetrahydrofuran is found under ETHERS. Polymers of tetrahydrofuran are covered under the general topic, POLYETHERS. Several other compounds containing the tetrahydrofuran ring, which are most readily produced from furfural, are discussed here.

The furan nucleus is a cyclic, dienic ether with some aromaticity (2). It is the least aromatic of the common 5-membered heterocycles. A comparison of the aromaticity (3) of several of these compounds is shown below.

Property	Furan	Thiophene	Pyrrrole	Benzene
Unified aromaticity index	53	81.5	85	100
Resonance energy, kJ mol	113.8	180.0	169.4	191.6
kcal mol	27.2	43.0	40.5	45.8

The balance between aromatic and aliphatic reactivity is affected by the type of substituents on the ring. Furan functions as a diene in the Diels-Alder reaction. With maleic anhydride, furan readily forms 7-oxabicyclo [2.2.1]hept-5-ene-2,3-dicarboxylic anhydride in excellent yield [5426-09-5] (4).



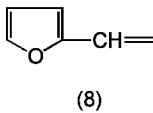
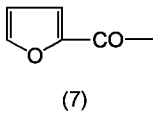
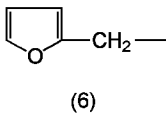
(4)

Alkylfurans, halofurans, alkoxyfurans, furfuryl ester and ethers, and furfural diacetate [613-75-2] behave similarly. Furans containing electron withdrawing constituents, for example, furfural, 2-furoic acid, and nitrofurans, fail as dienes even with very strong dienophiles.

Radicals derived from furan are named similarly to analogous radicals in the benzene series. Typical radicals are 2 (or α)-furyl (5), 2-furfuryl (6), 2-furoyl (7), and 2-furfurylidene (8):



(5)



Centers of high electron density occur at the *alpha* positions of the ring which governs the behavior of furans in electrophyllic substitution reactions. Such reactions proceed readily at available alpha positions. Orientation of incoming groups is similar to that observed in benzene compounds, eg, 3-methylfuran is nitrated to yield 2-nitro-3-methylfuran (similar to *ortho*) and 3-furancarboxaldehyde gives 2-nitro-4-furancarboxaldehyde (corresponding to *meta*). Furan compounds differ from benzene analogues in the ease in which they are polymerized; resinification or polymerization is easily brought about by strong acids. Resins or polymers are the principal end products of furfural utilization and account for a high percentage of the ultimate consumption of furfural.

For reviews on furan chemistry see References 1, 4–7. The monograph by Dunlop and Peters (1), published in 1953 remains as the most comprehensive review of furan chemistry available. Unfortunately, it has never been revised, has been out of print for many years, and is not readily available. An excellent, review by Dean (7) covers advances in furan chemistry up through about 1982.

Furfural

Furfural was first isolated in the early nineteenth century. Dobereiner is credited with the discovery. He obtained a small amount of a yellow "oil" (too little to characterize) as a by-product in the preparation of formic acid (8). Other chemists found that the same "oil" having a characteristic aroma could be obtained by boiling finely divided vegetable materials such as oats, corn, sawdust, bran, etc, with aqueous sulfuric acid or other acids (9,10). The oil was present in the liquid resulting from condensation of the vapors produced during heating. The empirical formula C₅H₄O₂ was determined by Stenhouse (10). Ring structure and location of the aldehyde group were established by the efforts of Baeyer, Markwald, and Harries (11–14). For some time, furfural was a laboratory curiosity and a compounding ingredient of perfumes.

It was not until the twentieth century that furfural became important commercially. The Quaker Oats Company, in the process of looking for new and better uses for oat hulls found that acid hydrolysis resulted in the formation of furfural, and was able to develop an economical process for isolation and purification. In 1922 Quaker announced the availability of several tons per month. The first large-scale application was as a solvent for the purification of wood rosin. Since then, a number of furfural plants have been built world-wide for the production of furfural and downstream products. Some plants produce as little as a few metric tons per year, the larger ones manufacture in excess of 20,000 metric tons.

Furfural can be classified as a reactive solvent. It resinifies in the presence of strong acid; the reaction is accelerated by heat. Furfural is an excellent solvent for many organic materials, especially resins and polymers. On catalyzation and curing of such a solution, a hard rigid matrix results, which does not soften on heating and is not affected by most solvents and corrosive chemicals.

Furfural is formed by a series of reactions when biomass materials containing hemicellulose are treated with acid at an elevated temperature. Hemicellulose is basically a polymer of 5-carbon sugars. Cellulose and starch are polymers of 6-carbon sugars. Most plant matter is made up of cellulose, hemicellulose, and lignin. When treated with aqueous acid, the hemicellulose is depolymerized to give primarily xylose, which under the reaction conditions loses three molecules of water and cyclizes to give furfural. The precise mechanism for the formation of furfural from C₅-sugars or their precursors has not yet been unequivocally established. Superheated steam passing through a reactor containing eg, ground oat hulls, provides heat for the conversion as well as carrying furfural away from the reaction mixture. Furfural also resinifies under the reaction conditions so the steam serves a very important function by removing the furfural as fast as it is formed. Most of the other compounds present in the reaction vessel are not volatile with steam. Condensation of the vapors gives a dilute solution of furfural in water. Furfural is isolated from the water by azeotropic distillation and phase separation, followed by dehydration and purification by distillation.

Physical Properties. Furfural [98-01-1] (2-furancarboxaldehyde), when freshly distilled, is a colorless liquid with a pungent, aromatic odor reminiscent of almonds. It darkens appreciably on exposure to air or on extended storage. Furfural is miscible with most of the common organic solvents, but only slightly miscible with saturated aliphatic hydrocarbons. Inorganic compounds, generally, are quite insoluble in furfural.

The most important physical properties of furfural, as well as similar properties for furfuryl alcohol, tetrahydrofurfuryl alcohol and furan are given in Table 1. The tabulated properties of furfural are supplemented by a plot (Fig. 1) of the vapor–liquid compositions for the system, furfural–water (15,16).

Table 1. Physical Properties of Furan Derivatives

	Furfural	Furfuryl alcohol	Furan	Tetrahydrofurfuryl alcohol
<i>General Properties</i>				
molecular weight	96.09	98.10	68.08	102.13
boiling point at 101.3 kPa (1 atm), °C	161.7	170	31.36	178
freezing point, °C	36.5		85.6	< 80
metastable crystalline form		29		
stable crystalline form		14.63		
refractive index, <i>n</i> _D				
20°C	1.5261	1.4868	1.4214	1.4250
25°C	1.5235			1.4499
density, <i>d</i> ₄ , at 20°C	1.1598	1.1285	0.9378	1.0511
vapor pressure, 100 Pa (0.75 mm Hg)				
15°C			130	
0°C			277	
20°C			658	
50°C			1980	
60°C	21	8.5		10

80°C	56	25	28	
100°C	132	58	68	
120°C	280	164	156	
140°C	567	361	324	
vapor density (air = 1)	3.3	3.4	2.36	3.5
critical pressure, P_c , MPa ^a	5.502		5.32	
critical temperature, T_c , °C	397		214	
solubility, wt %, in water				
20°C	8.3	∞		∞
25°C			1	
alcohol; ether	∞	∞	∞	∞
	<i>Thermodynamic properties</i>			
heat of vaporization, kJ mol ^b	38.6		27.1	
specific heat (liq), J (gK) ^b				
20°C			1.699	
20–27°C				1.774
20–100°C	1.741			
25°C		2.100		
specific heat (vap), J (gK) ^b				
31.36°C			1.021	
98.99°C			1.251	
heat of combustion (liq), kJ mol ^b	2344	2548	2092	2965
	<i>Fluid properties</i>			
viscosity, mPas (cP)				
20°C			0.38	6.24
25°C	1.49	4.62		
surface tension, mN m (dyn cm)				
25°C		ca 38		37
29.9°C	40.7			
	<i>Electrical properties</i>			
dielectric constant				
20°C	41.9			
23°C				13.6
	<i>Flammability properties</i>			
explosion limits (in air), vol %	2.1–19.3	1.8–16.3	2.3–14.3	1.5–9.7
flash point, °C				
Tag closed cup	61.7	65	35.5	
Tag open cup				83.9
ignition temperature, °C	315	391		282

^a To convert MPa to atm, divide by 0.101.
^b To convert J to cal, divide by 4.184.

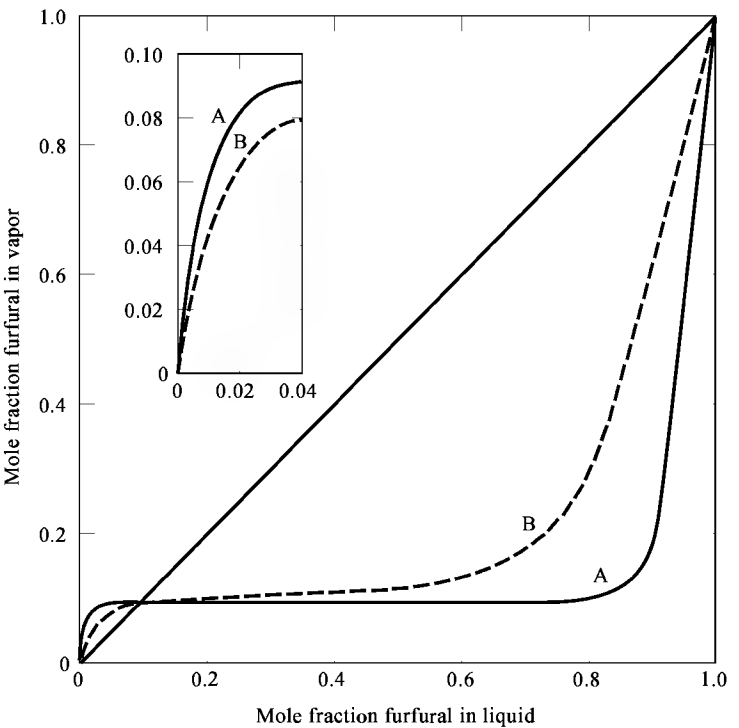


Fig. 1. Vapor–liquid equilibria in the furfural–water system. A at 101 kPa (1 atm) (15). B at 598 kPa (5.92 atm) (16).

Chemical Properties. The chemical properties of furfural are generally characteristic of aromatic aldehydes but with some differences attributable to the furan ring. Furfural resinifies in the presence of acid and heat. Open chain compounds are formed from furfural under strong oxidizing conditions. Because furfural has a fairly low degree of aromaticity, the ring is more easily saturated than the benzene ring on catalytic hydrogenation.

Depending on catalysts and conditions, products of hydrogenation can include furfuryl alcohol [98-00-0], tetrahydrofurfuryl alcohol [97-99-4], 2-methylfuran [534-22-5] and even 2-methyltetrahydrofuran [96-47-9]. Under strongly reducing conditions, the ring is opened.

Furfural is very thermally stable in the absence of oxygen. At temperatures as high as 230°C, exposure for many hours is required to produce detectable changes in the physical properties of furfural, with the exception of color (17). However, accelerating rate calorimetric data shows that a temperature above 250°C, in a closed system, furfural will spontaneously and exothermically decompose to furan and carbon monoxide with a substantial increase in pressure. The pressure may increase to 5000 psi or more, sufficient to shatter the container (18).

Furfural can be oxidized to 2-furoic acid [88-14-2], reduced to 2-furanmethanol [98-00-0], referred to herein as furfuryl alcohol, or converted to furan by decarbonylation over selected catalysts. With concentrated sodium hydroxide, furfural undergoes the Cannizzaro reaction yielding both 2-furfuryl alcohol and sodium 2-furoate [57273-36-6].

Acetals are readily formed with alcohols and cyclic acetals with 1,2 and 1,3-diols (19). Furfural reacts with poly(vinyl alcohol) under acid catalysis to effect acetalization of the hydroxyl groups (20,21). Reaction with acetic anhydride under appropriate conditions gives the acylal, furfurylidene diacetate [613-75-2] (22,23).

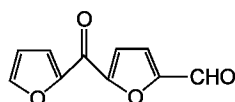
Nitration and halogenation of furfural occurs under carefully controlled conditions with introduction of the substituent at the open 5-position (24,25). Nitration of furfural is usually carried out in the presence of acetic anhydride, resulting in the stable compound, 5-nitrofurfurylidene diacetate (26,27). The free aldehyde is isolated by hydrolysis and must be used immediately in a reaction because it is not very stable.

Just as most other aldehydes do, furfural condenses with compounds possessing active methylene groups such as aliphatic carboxylic esters and anhydrides, ketones, aldehydes, nitriles, and nitroparaffins.

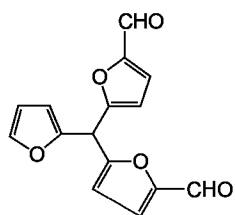
The electron-withdrawing character of the aldehyde group makes furfural quite resistant to hydrolytic fission. Even at high temperatures, long exposure is required to effect extensive ring opening destruction of furfural by dilute acids (28). Under these conditions a competing reaction takes place: black polymer results, and the rate of its formation is dependent upon the hydrogen ion concentration and temperature.

Furfural is a resin former under the influence of strong acid. It will self-resinify as well as form copolymer resins with furfuryl alcohol, phenolic compounds, or convertible resins of these. Conditions of polymerization, whether aqueous or anhydrous, inert or oxygen atmosphere, all affect the composition of the polymer. Numerous patents have issued relating to polymerization and to applications. Although the resins exhibit a degree of brittleness, they have many outstanding properties; a number of applications are discussed under "Uses."

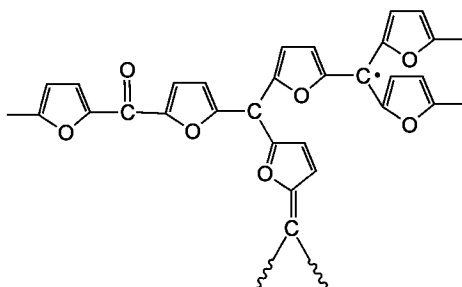
Several early interpretations of the polymerization mechanism have been proposed (1,17,29–31). Because of the complexity of this polymerization and insoluble character of the products, key intermediates have not ordinarily been isolated, nor have the products been characterized. Later work, however, on the resinification of furfural (32,33) has provided a new insight on the polymerization mechanism, particularly with respect to thermal reaction at 100–250°C in the absence of air. Based on the isolation and characterization of two intermediate products (9) and (10), structure (11) was proposed for the final resin. This work also explains the color produced during resinification, which always is a characteristic of the final polymer (33). The resinification chemistry is discussed in a recent review (5).



(9) [57583-51-4]



(10) [72918-23-1]



(11)

The presence of stable free radicals in the final polycondensate is supported by the observation that traces of (11) have a strong inhibiting effect on the thermal polymerization of a number of vinyl monomers. Radical polymerization was inhibited to a larger extent by a furfural resin than by typical polymerization inhibitors (34). Thermal degradative methods have been used to study the structure of furfural resinified to an insoluble and infusible state, leading to proposed structural features (35).

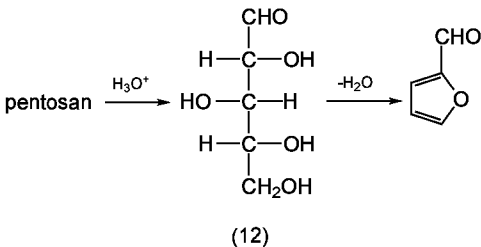
Manufacture. Furfural is produced from annually renewable agricultural sources such as nonfood residues of food crops and wood wastes. The pentosan polysaccharides, xylan and arabinan, commonly known as hemicellulose, are the principal precursors of furfural and are always found together with lignin and cellulose in plant materials.

By-products such as corn cobs, cottonseed hulls, oat hulls, rice hulls, and cereal grasses constitute one large natural storehouse; bagasse, a by-product of sugar-cane harvesting, another; and wood and wood products yet another. The chief constituent of the pentosan fraction of these materials is xylan, a polysaccharide having backbone chains of β -D-xylopyranosyl residues (33). According to published data (34), xylan with a small amount of arabinan, a highly branched polysaccharide of α -arabinofuranosyl residues (35), accounts for 2–30% of cereal straws and grains, approximately 15–25% of deciduous woods and 5–15% of coniferous woods, based on the dry matter of these plants.

Theoretically, all pentosan-containing substances are potentially usable for the production of furfural. Only a relatively few, however, are commercially significant (1,36,37). Corn cobs, oat hulls, rice hulls, and bagasse contain sufficient levels of pentosans and are often easily available in large tonnages within a limited radius of a furfural-producing plant. The simultaneous occurrence of side reactions complicates the recovery of furfural and limits its practical yield, which is far from quantitative under any commercial mode of operation. The average weight yield of furfural based on the dry weight of

these raw materials is 10° or less. Thus, the economics of furfural production is highly dependent upon pentosan content, availability, and cost of raw materials, as well as the expense of collecting, shipping, and handling them.

Furfural is commercially produced in batch or continuous digesters (37) where the pentosans are first hydrolyzed to pentoses (primarily xylose (12), which are then subsequently cyclodehydrated to furfural:



Strong inorganic acid can be used as a reaction catalyst, or if the temperature is raised high enough, sufficient acetic acid is produced by the heat to catalyze formation of furfural.

In all processes, raw material is charged to the digester and heated with high pressure steam. Enough excess steam is used to drive the furfural out of the reaction zone as vapor. The condensed reactor vapors are fed to a stripping column from which an enriched furfural–water distillate mixture is taken overhead and condensed. The liquid passes into a decanter where it separates into two layers. The furfural-rich lower layer containing about 6° water is processed further to obtain the furfural of commerce, and the water-rich layer containing about 8° furfural is recycled back to the stripper column as reflux. The last process steps are dehydration and distillation. The furfural isolation and recovery process is illustrated in Figure 2. Flow charts for the Escher-Wyss and Rosenlew, two slightly different processes can be found in Ref. 38. In general, regardless of the process employed, the cellulose in the raw material undergoes only partial degradation and remains with the lignin in the by-product residue.

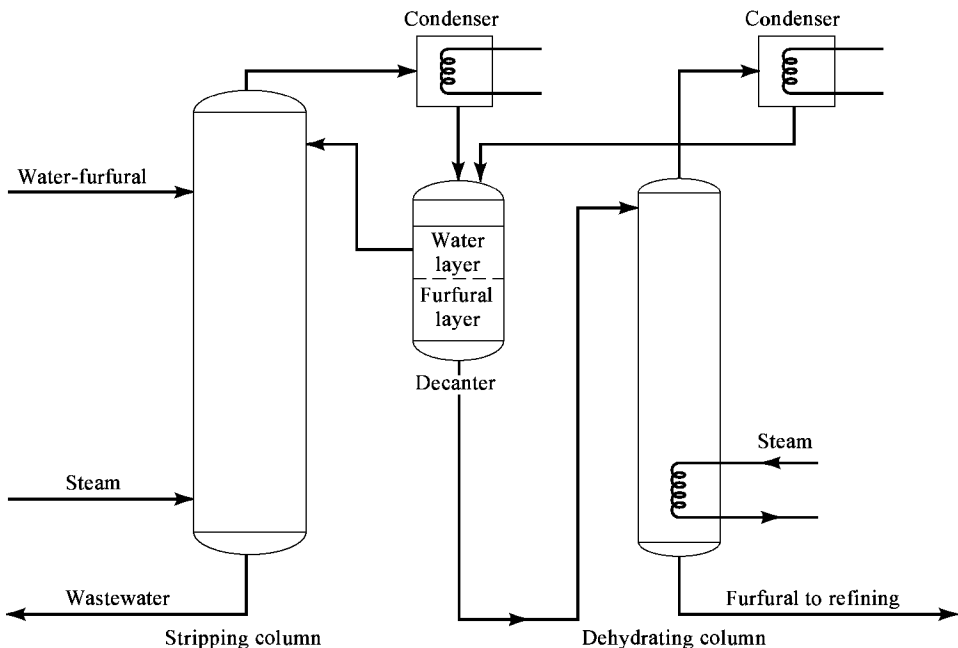


Fig. 2. Furfural recovery from aqueous solutions.

Most laboratories now use gas chromatographic methods for the analysis of refined furfural. For dilute aqueous solutions, wet methods are sometimes more appropriate. A number of wet analytical methods for quantitative estimation of furfural have been reported (1), yet none have universal acceptance. For the most part, they can be divided into two general categories: those involving reactions of the aldehyde group, and those that are dependent upon the unsaturated characteristics of the furan ring. Accordingly, the determination of furfural is generally subject to interference by carbonyl compounds or other unsaturated substances. The quantitative measurement of furfural in water may be carried out by uv spectroscopy; absorbance is measured at 276 nm (39). This method is best used on dilute solutions. However, uv-absorbing impurities such as phenolic compounds interfere with the analysis. The Hughes-Acree method, based on the reaction of bromine with the double bonds in furfural was at one time the method of choice (1).

The estimation of furfural potential of various raw materials is best done by the AOAC method (1). Although liquid chromatographic methods are now available for the estimation of polymeric pentosans, results do not always correlate well with furfural formation.

Furfural is not corrosive to metal and can be shipped in mild-steel tank cars or trucks or steel drums. Storage in either aboveground or underground installations is satisfactory. For extended storage with maximum stability, cool storage conditions and nitrogen blanketing are recommended. Because furfural is an excellent solvent and penetrant, care must be taken that all joints are secure and that the pump and valve packings are in good condition. Unopened drums may be stored in cool locations for months without appreciable change in physical properties.

The flash point of furfural is 143°F by Tag Closed Cup. Because of its chemical reactivity, furfural should be kept away from strong acids, alkalis or strong oxidizing chemicals. When furfural is stored for long periods in contact with air, there is a gradual darkening of color, increase in acidity, and formation of a soluble polymer.

According to the latest available information (40), in 1992, total world-wide capacity for furfural was 240,000 metric tons (t) (530,000,000 lb). Production capability is moving away from developed countries into developing nations. The Peoples' Republic of China has become a significant factor in furfural production, with an estimated total capacity of 50,000 t, most of it being produced in small plants. New capacity has come on stream or is being planned in other Asian countries as well as in Indonesia. South African capacity is rated at 17,000 t. United States capacity was estimated at 73,000 t (30° of the world total), production at 40,000 t and consumption at 43,000. Western Europe capacity was 8,000 t (down from 27,000 in 1988). Production was 6,000 t and consumption was 36,000 t, a significant amount being imported from the western hemisphere. Japan is not presently a producer of furfural, importing approximately 9,000 t, mostly from China. At this time there is significant overcapacity worldwide; consumption is steady or declining slightly. The United States list price for furfural was \$0.79 /lb in 1992; it has not changed significantly since then.

Uses. Furfural is primarily a chemical feedstock for a number of monomeric compounds and resins. One route produces furan by decarbonylation. Tetrahydrofuran is derived from furan by hydrogenation. Polytetramethylene ether glycol [25190-06-1] is manufactured from tetrahydrofuran by a ring opening polymerization reaction. Another route (hydrogenation) produces furfuryl alcohol, tetrahydrofurfuryl alcohol, 2-methylfuran, and 2-methyltetrahydrofuran. A variety of proprietary synthetic resins are manufactured from furfural and or furfuryl alcohol. Other

chemicals that have been derived, on a lesser scale are described in a later section. In this country more than 90% of the furfural produced is used as feedstock for other compounds. A breakdown of U.S. consumption in 1992 is shown in Table 2.

Table 2. U.S. Consumption of Furfural

Use	t	(10 lb)
raw material for furfuryl alcohol	24,460	(54)
raw material for tetrahydrofurfuryl alcohol	905	(2)
raw material for tetrahydrofuran	13,590	(30)
solvent applications	3,625	(8)
other uses	1,815	(4)

As can be seen, most of the furfural produced in this country is consumed as an intermediate for other chemicals. Hydrogenation to furfuryl alcohol is the largest use. Some of the furfuryl alcohol is further hydrogenated to produce tetrahydrofurfuryl alcohol. The next major product is furan, produced by decarbonylation. Furan is a chemical intermediate, most of it is hydrogenated to tetrahydrofuran, which in turn is polymerized to produce polytetramethylene ether glycol (PTMEG).

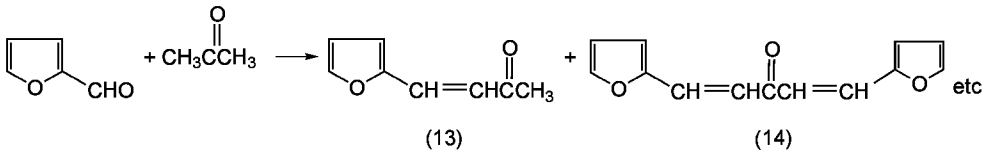
Furfural and furfuryl alcohol are specialty solvents. They also are reactive solvents and contribute low viscosity to resin formulations. Thermosetting resins containing furfural and or furfuryl alcohol. "Furan resins" demonstrate specialty properties including corrosion resistance, high carbon yield, stability at elevated temperature, low fire hazard, and excellent physical strength. These properties are of industrial importance in making foundry molds and cores, fiberglass composites, mortars, cements, plastic insulation foams, refractory mixes, high carbon composites, and aggregate binders, among others.

The principal direct application of furfural is as a selective solvent. It is used for separating saturated from unsaturated compounds in petroleum refining, for the extractive distillation of butadiene and other C₄ hydrocarbons in the manufacture of synthetic rubber; and for the production of light-colored wood rosin. These applications for furfural are mature and volume is declining. Furfural is both a solvent and a processing aid for refining anthracene. It is an ingredient in furfuryl alcohol resins and in phenol–aldehyde resins. Furfural is also used as a resin solvent and wetting agent in the manufacture of abrasive wheels and brake linings. Furfural is used as a reactive solvent in cold-blending of pitch. The mixture gives resinous impregnants that interact at carbonization temperatures. These resin pitch combinations have curing and carbonization characteristics better than those of either component alone. The cold-blending process significantly improves the environmental acceptance of pitch and offers industry an alternative to hot processing.

High quality carbon fibers have been produced (41) by pyrolyzing, in a nonoxidizing atmosphere, resinous fibers prepared from a viscous copolymer of furfural and pyrrole. Yield of carbon fiber having a glossy surface and only 3% voids was 66%. Compared with other graphite-type fibers, these carbon fibers had improved mechanical strength, higher oxidation resistance, and lower density.

Useful thermosetting resins are obtained by interaction of furfural with phenol. The reaction occurs under both acidic and basic catalysis. Other large uses of furfural together with phenol are in the manufacture of resin-bonded grinding wheels and coated abrasives (5).

Furfural reacts with ketones to form strong, crosslinked resins of technical interest in the former Soviet Union; the U.S. Air Force has also shown some interest (42,43). The so-called furfurylidene acetone monomer, a mixture of 2-furfurylidene methyl ketone [623-15-4], (13), bis-(2-furfurylidene) ketone [886-77-1] (14), mesityl oxide, and other oligomers, is obtained by condensation of furfural and acetone under basic conditions (44,45). Treatment of the "monomer" with an acidic catalyst leads initially to polymer of low molecular weight and ultimately to cross-linked, black, insoluble, heat-resistant resin (46).



Furfural–acetone resins have been used to form resin-aggregate mixtures referred to as organic concretes. Despite the reportedly excellent properties, there has been virtually no commercial use of such resins outside the former Soviet Union. The structures and polymerization mechanisms of these furfural–aldehyde–ketone polymers are discussed in a review (6).

Furfural has been used as a component in many resin applications, most of them thermosetting. A comprehensive review of the patent literature describing these uses is beyond the scope of this review. A few, selected recent patents and journal articles have been referenced. Resins prepared from the condensation products of furfural with urea (47), formaldehyde (48), phenols (49,50), etc, modified by appropriate binders and fillers are described in the technical literature; for earlier applications, see reference 1, which contains many references in an appendix.

Furfuryl Alcohol

Physical Properties. Furfuryl alcohol (2-furanmethanol) [98-00-0] is a liquid, colorless, primary alcohol with a mild odor. On exposure to air, it gradually darkens in color. Furfuryl alcohol is completely miscible with water, alcohol, ether, acetone, and ethyl acetate, and most other organic solvents with the exception of paraffinic hydrocarbons. It is an excellent, highly polar solvent, and dissolves many resins.

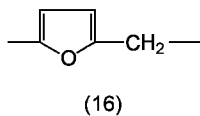
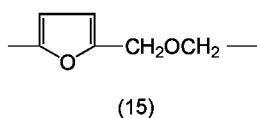
The physical constants of furfuryl alcohol are listed in Table 1. When exposed to heat, acid or air the density and refractive index of furfuryl alcohol changes owing to chemical reaction (51), and the rate of change in these properties is a function of temperature and time of exposure.

Chemical Properties. The chemical properties of furfuryl alcohol are characteristic of the hydroxymethyl group; however, in the presence of acid the reactivity of the furan ring is significantly enhanced through allyl carbonium ion formation. Furfuryl alcohol undergoes the typical reactions of a primary alcohol such as oxidation, esterification, and etherification. Although stable to strong alkali, furfuryl alcohol is very sensitive to acid, thus imposing limitations on the conditions used in many of the typical alcohol reactions. For example, esterifications must be performed under neutral or basic conditions, or via base-catalyzed transesterification. Likewise, etherifications are done under basic condensation conditions to suppress polymerization of the furan ring. Yet it is this very sensitivity to acid that is the basis for most of the commercial uses of furfuryl alcohol. It is a more reactive solvent than furfural and easily resinifies under the influence of acid and heat.

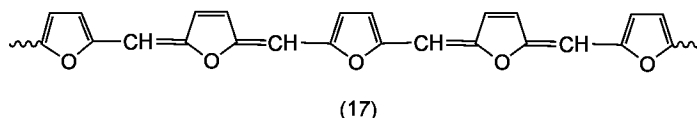
Under acidic conditions, furfuryl alcohol polymerizes to black polymers, which eventually become crosslinked and insoluble in the reaction medium. The reaction can be very violent and extreme care must be taken when furfuryl alcohol is mixed with any strong Lewis acid or Brønstad acid. Copolymer resins are formed with phenolic compounds, formaldehyde and or other aldehydes. In dilute aqueous acid, the predominant reaction is a ring opening hydrolysis to form levulinic acid [123-76-2] (52). In acidic alcoholic media, levulinic esters are formed. The mechanism for this unusual reaction in which the hydroxymethyl group of furfuryl alcohol is converted to the terminal methyl group of levulinic acid has recently been elucidated (53).

Hydrogenation of furfuryl alcohol can yield 2-tetrahydrofurfuryl alcohol, 2-methylfuran, 2-methyltetrahydrofuran, or straight-chain compounds by hydrogenolysis of the ring. Ethoxylation and propoxylation of furfuryl alcohol provide useful ether alcohols.

The chemistry of furfuryl alcohol polymerization has received much attention over the years. Several recent reviews have been written (5,6,54). Based on the accumulated data, furfuryl alcohol has to be considered a bifunctional monomer in the initial stage and its "normal" reactions give linear chains or oligomers containing essentially two repeating units (15,16) with (16) predominating.



In 1996, the mechanism of resinification was clearly and elegantly established (55). Development of color (a phenomenon which up until now has never been adequately explained) has been shown to be due to conjugated sequences along linear oligomeric chains (17). Branching and crosslinking, which cause gelation and insolubilization are described to take place through the reaction of terminal methylol groups or disubstituted furan rings with nonfuranic double bonds in the oligomeric chain.



Manufacture. Furfuryl alcohol has been manufactured on an industrial scale by employing both liquid-phase and vapor-phase hydrogenation of furfural (56,57). Copper-based catalysts are preferred because they are selective and do not promote hydrogenation of the ring.

Furfuryl alcohol is shipped in bulk or drums. Although not corrosive to metals, it is a powerful solvent and penetrant; containers, tanks, lines, and valves need to be in good condition to avoid potential leakage. Furfuryl alcohol can be stored in containers lined with baked phenolic resin coatings; however, it should not be put in containers that are coated with lacquers, varnishes, or epoxy resins because it is an excellent solvent for many such coatings.

Furfuryl alcohol, on long storage, becomes progressively darker and less water soluble, a change that is also caused by heat, acidity, and exposure to air. The reactions responsible for this change in water solubility may be retarded by the addition in small quantity of an organic or inorganic base. Commercial furfuryl alcohol, however, usually does not contain any additives.

Furfuryl alcohol is comparable to kerosene or No. 1 fuel oil in flammability, the Tag Closed Cup flash point is 170°F. In the presence of concentrated mineral acids or strong organic acids, furfuryl alcohol reacts with explosive violence. Therefore, precautions should be taken to avoid contact of such materials with the alcohol. Caution is also recommended to avoid over-catalysis in the manufacture of furfuryl alcohol resins.

Worldwide furfuryl alcohol capacity in 1993 was estimated to be 110,000 metric tons (38). As with furfural, new capacity in developing countries is replacing older capacity in developed countries. China and South Africa have become significant producers of furfuryl alcohol. New plants have been built in Asia and Indonesia as well. Consumption of furfuryl alcohol is spread over the globe; the largest use is in the foundry industry which is increasingly moving away from heavily industrialized countries.

Uses. Furfuryl alcohol is widely used as a monomer in manufacturing furfuryl alcohol resins, and as a reactive solvent in a variety of synthetic resins and applications. Resins derived from furfuryl alcohol are the most important application for furfuryl alcohol in both utility and volume. The final cross-linked products display outstanding chemical, thermal, and mechanical properties. They are also heat-stable and remarkably resistant to acids, alkalies, and solvents. Many commercial resins of various compositions and properties have been prepared by polymerization of furfuryl alcohol and other co-reactants such as furfural, formaldehyde, glyoxal, resorcinol, phenolic compounds and urea. In 1992, domestic furfuryl alcohol consumption was estimated at 47 million pounds (38).

The polymerization or resinification of furfuryl alcohol is highly exothermic, and requires careful control of the reaction conditions. Proper attention to safety cannot be overemphasized. The addition of too much strong acid will lead to a violent, almost explosive, reaction. The resinification of furfuryl alcohol is autocatalytic in the sense that the rate increases geometrically with temperature. During a resinification reaction, temperature control is generally accomplished with refluxing solvent and or external cooling. As a safety precaution, to prevent a runaway reaction, provision for emergency neutralization of the catalyst is essential. At temperatures above 250°C, in a closed system, furfuryl alcohol undergoes a strong exothermic reaction which is not controlled by base (18). Furfuryl alcohol should not be heated to above 250°C in a closed system.

In practice, intermediate, liquid resins, capable of further reaction are usually prepared. Polymerization is carried to an established end-point as determined by viscosity or other measurements. When the proper end-point has been reached, the reaction is terminated by adjusting the pH of the system to 5–8. Such liquid resins can be stored for six months or longer, then catalyzed and reacted further to obtain the final, desired product.

Furfuryl alcohol is an excellent solvent for many resins. A number of applications are based on its reactive solvent properties. When a resin solution is cured, a hard, rigid, thermoset matrix results. The final, cured product often has many outstanding properties. For example, furfuryl alcohol is a reactive solvent for phenolic resins in the manufacture of refractories for ladles holding molten steel. The furfuryl alcohol provides a low viscosity to allow good mixing of the resin with the refractory particles, then reacts with the phenolic resin during curing. When heated to high temperatures, the matrix carbonizes, producing a strong refractory bond.

The industrial value of furfuryl alcohol is a consequence of its low viscosity, high reactivity, and the outstanding chemical, mechanical, and thermal properties of its polymers, corrosion resistance, nonburning, low smoke emission, and excellent char formation. The reactivity profile of furfuryl alcohol and resins is such that final curing can take place at ambient temperature with strong acids or at elevated temperature with latent acids. Major markets for furfuryl alcohol resins include the production of cores and molds for casting metals, corrosion-resistant fiber-reinforced plastics (FRPs), binders for refractories and corrosion-resistant cements and mortars.

Metal-Casting Cores and Molds. In the foundry industry there is a highly significant binder technology based on furfuryl alcohol resins. These binders are cured at ambient temperature or at elevated temperatures. Microwave curing has been used. Even curing with acidic gases is feasible. Total furfuryl alcohol based binder consumption by the foundry industry in 1992 has been estimated at 35 million pounds, representing 27 million pounds of contained furfuryl alcohol (38).

Furan No-Bake (FNB) resins are used worldwide in competition with several other chemical binder systems. FNB binders are used in the production of sand cores and molds that cure without heat (58–61). The FNB process uses a reactive furan resin binder with strong organic acid or phosphoric acid as catalyst. Based on the requirements of the particular system, slow or rapid curing can be accomplished at ambient temperatures, eliminating the need for long ovenbaking cycles. Accelerators are sometimes used to increase curing rate (62–65). The catalyst is added to the sand, followed by the binder, often in a continuous mixer. The characteristics of the acid and the binder as well as the mixing ratio determine curing rate characteristics. Other money saving benefits that accrue include: fewer or no reinforcing rods; less pattern wear; dimensional accuracy of the castings; easy shakeout of core sand after pouring; and minimal clean-up of the finished casting. The used sand can easily be reclaimed by mechanical means. Most important, the process is simple, requiring fewer steps to produce cores, and can be operated by semiskilled personnel.

Furan hot-box resins are used in both ferrous and nonferrous foundries (66,67). In this process, resin and catalyst are intimately mixed with dry sand and then blown into heated metal boxes containing a cavity the shape of the desired core. In seconds, the surface of the sand mass hardens and, as soon as the core has cured sufficiently to be rigid and handleable the box is opened and the core removed. Automotive cores with excellent dimensional accuracy and high strengths are made via this forty-year-old process.

Both FNB and Hot Box applications are mature and declining as new technology is being used more and more in the foundry industry.

Technological advances continue to be made, several recent patents describe advanced phenol–formaldehyde–furfuryl alcohol binder systems (68–70). These systems are free of nitrogen compounds that can be detrimental to metal integrity. Systems with extended bench life have also been developed (71).

One newer application, however, is not yet mature. Warm Box systems, curing at significantly lower temperatures, are becoming more widely used. The warm-box resin is based on 2,5-bis(hydroxymethyl)furan (BHMF) resins. Lower binder levels are reported effective at core-curing temperatures of 149–177°C (72). The binders do not contain any nitrogen compounds. An essential part of the new technology is new latent curing catalysts (73,74). Core production rate using warm-box resins is equivalent to the rate with hot-box resins and provides significant improvement in working environment. Lower emissions, both at core-making and pour off are claimed.

Corrosion Resistant Fiber-Reinforced Plastic (FRP). Fiber glass reinforcement bonded with furfuryl alcohol thermosetting resins provides plastics with unique properties. Excellent resistance to corrosion and heat distortion coupled with low flame spread and low smoke emission are characteristics that make them valuable as laminating resins with fiber glass (75,76). Another valuable property of furan FRP is its strength at elevated temperature. Hand-layup, spray-up, and filament-winding techniques are employed to produce an array of corrosion-resistant equipment, pipes, tanks, vats, ducts, scrubbers, stacks, and reaction vessels for industrial applications throughout the world.

Reinforced furan resins have been used for many years in process piping and in underground sewer or waste-disposal systems. With a wide range in pH acceptability and good solvent resistance, furan piping has been a logical choice for many services.

The highly cross-linked thermoset character of cured furan resin provides substantially greater corrosion resistance than other resins used in fiber glass-reinforced plastics. Hence, furan FRP is recommended (77) for high performance chemical process equipment in service with chlorinated aromatics, highly oxygenated organic solvents, and other admixtures of organics and aggressive aqueous media of a nonoxidizing nature. Commercially manufactured corrosion-resistant off-the-shelf pipe using furan resins and fiber glass has been available since 1977 (77).

The inherent fire resistance and low smoke properties of furan resins appear to be related to the high degree of charring that takes place with minimum evolution of volatiles when exposed to fire.

Cements and Mortars. One of the original applications for furan resins was in the jointing of bricks and masonry. Furan resins have been extensively used for formulating mortars, grouts, and "setting beds" for brick linings in structures exposed to corrosive media, especially concentrated acids, and for setting tile in floors exposed to alkaline cleaning solutions or other corrosive chemical media. Mortars and grouts for these applications are usually formulated as two-package systems and mixed just prior to use. Silica fillers are satisfactory for many applications, but carbon flour (ground coke or other nongraphitic, nonactivated carbon) is used where resistance to hydrogen fluoride, fluoride salts, or hot, concentrated alkaline cleaning solutions is required.

High Carbon Yield. Furfuryl alcohol and furfural are reactive solvents (monomers) and are effective in producing high carbon yield (heat induced carbonization in a reducing atmosphere). They function as binders for refractory materials or carbon bodies. Furfuryl alcohol usually requires acidic catalysis and furfural basic catalysis. Mixtures of furfuryl alcohol and furfural are generally catalyzed with acid although some systems may be catalyzed with base.

Furfuryl alcohol alone, or in combination with other cross-linkable binders such as phenolic resins, chemical by-products and pitch, catalyzed with acid, gives carbon yields of 35–56%. Furfural together with cyclohexanone, pitch, or phenolic resins gives, under acid catalysis, yields of 35–55% carbon; under basic catalysis yields of 5–50% are achieved. Furfurylideneacetone resins (13 and 14), catalyzed by acid or base, give carbon yields of 48–56 and 30–35%, respectively (78).

Other Uses. Anisotropic and isotropic carbon are produced from furfural-modified systems; glassy carbon is produced primarily from furfuryl alcohol or BHMF resins (78,79).

Furfuryl alcohol is used alone or in combination with other solvents for various cleaning and paint removing operations. The ethylene oxide adduct of furfuryl alcohol is especially useful in this type of application (80–83).

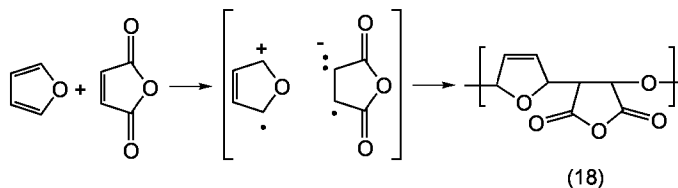
Furan

Physical Properties. Furan, a colorless liquid with a strong ethereal odor, is low-boiling and highly flammable. It is miscible with most common organic solvents but only very slightly soluble in water. The physical properties of furan are listed in Table 1.

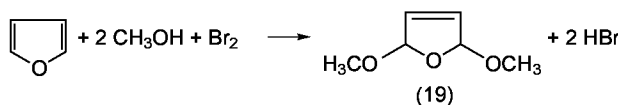
Chemical Properties. Furan is a heat-stable compound, although at 670°C in the absence of catalyst, or at 360°C in the presence of nickel, it decomposes to form a mixture consisting mainly of carbon monoxide, hydrogen, and hydrocarbons (84). Substitution and addition reactions can be effected under controlled conditions, with reaction occurring first in the 2- and 5-positions. Gas-phase reaction with ammonia over a modified silica–alumina catalysts gives pyrrole in good yield, 80% conversion with 72% selectivity (85). Carrying out the reaction in steam using HZSM-5 catalyst gives 65% conversion with 99% selectivity (86). The reaction of furan with primary amines gives *N*-substituted pyrroles although in low yield (87). Acylation under Friedel-Crafts conditions goes readily.

Furan and maleic anhydride undergo the Diels-Alder reaction to form the tricyclic 1:1 adduct, 7-oxabicyclo [2.2.1]hept-5-ene-2,3-dicarboxylic anhydride (4) in excellent yield. Other strong dienophiles also add to furan (88). Although both endo and exo isomers are formed initially, the former rapidly isomerize to the latter in solution, even at room temperature. The existence of a charge-transfer complex in the system has been demonstrated (89,90).

The radical-catalyzed polymerization of furan and maleic anhydride has been reported to yield a 1:1 furan-maleic anhydride copolymer (89,91). The structure of the equimolar product, as shown by nmr analyses, is that of an unsaturated alternating copolymer (18) arising through homopolymerization of the intermediate excited donor–acceptor complex (91,92).



Furan can be catalytically oxidized in the vapor phase with oxygen-containing gases to maleic anhydride (93). Oxidation with bromine or in an electrochemical process using bromide ion gives 2,5-dimethoxy-2,5-dihydrofuran [332-77-4] (19) which is a cyclic acetal of maleic dialdehyde (94–96).



Catalytic hydrogenation of furan to tetrahydrofuran is accomplished in either liquid or vapor phase. Hydrogenation of the double bonds is essentially quantitative over nickel catalysts but is generally accompanied by hydrogenolysis over the noble metals.

Manufacture. Furan is produced commercially by decarbonylation of furfural in the presence of a noble metal catalyst (97–100). Nickel or cobalt catalysts have also been reported (101–103) as well as noncatalytic pyrolysis at high temperature. Furan can also be prepared by decarboxylation of 2-furoic acid; this method is usually considered a laboratory procedure.

Furan, a Class 3 DOT hazard, is shipped in tank cars, tank trucks and in heavy duty drums. The selling price of furan in bulk is listed at \$1.60/lb.

Furan should be kept from heat and flame because of its low boiling point, low flash point, and high flammability. Unstabilized furan slowly forms an unstable peroxide on exposure to air and, therefore, care should be taken when using furan. When distilling furan, remove peroxides first by chemical

reaction, and never distill to dryness.

Uses. Furan is utilized as a chemical building block in the production of other industrial chemicals for use as pharmaceuticals, herbicides, stabilizers, and fine chemicals. There are a great many references to the use of furan as an intermediate in these applications. For a recent review, see Reference 104. Several of the principal uses are described below.

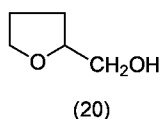
Furan is readily hydrogenated, hence it is a source of commercial tetrahydrofuran (THF). Reaction with hydrogen sulfide over alumina produces thiophene. A recent patent describes a catalyst based on cobalt and molybdenum oxides supported on alumina for this reaction (105). Furan undergoes the Diels-Alder reaction with strong dienophiles. Hydrogenation of the product resulting from reaction with maleic anhydride, followed by hydrolysis and neutralization gives a herbicide, Endothall.

The oxidative reaction of furan with bromine in methanol solution or an electrochemical process using sodium bromide produces 2,5-dimethoxy-2,5-dihydrofuran (19), which is a cyclic acetal of maleic dialdehyde. The double bond in (19) can be easily hydrogenated to produce the corresponding succindialdehyde derivative. Both products find application in photography and as embalming materials, as well as other uses.

Washing and cleaning agents containing salts of maleic acid–furan copolymers (106) form complexes with alkaline-earth ions. These cleaning compositions do not contain phosphorus or nitrogen and find use in metal, foodstuff, and machine dishwashing products.

Tetrahydrofurfuryl Alcohol

Physical Properties. Tetrahydrofurfuryl alcohol (2-tetrahydrofuranmethanol) [97-99-4] (20) is a colorless, high-boiling liquid with a mild, pleasant odor. It is completely miscible with water and common organic solvents. Tetrahydrofurfuryl alcohol is an excellent solvent, moderately hydrogen-bonded, essentially nontoxic, biodegradable, and has a low photochemical oxidation potential. Most applications make use of its high solvency. The more important physical properties of tetrahydrofurfuryl alcohol are listed in Table 1.



Chemical Properties. Without inhibitors, tetrahydrofurfuryl alcohol is susceptible to autoxidation, developing color and carbonyl functionality. In the absence of air, however, no observable changes occur even after several years storage. In the presence of air, if a stabilizer such as Naugard is added, tetrahydrofurfuryl alcohol remains colorless after protracted periods of storage. Peroxide accumulation is low, not dangerous, and readily dischargeable on redistillation.

The reactions of tetrahydrofurfuryl alcohol are characteristic of its structure, involving primary alcohol and cyclic ether functional groups. As a primary alcohol, it undergoes normal displacement or condensation reactions affording new functional groups, (eg, halides, esters, alkoxylates, ethers, glycidyl ethers, cyanoethyl ethers, amines, etc). As a cyclic ether, it is typically unreactive, but the ring can be forced to open by hydrolysis or hydrogenolysis to give a variety of open-chain compounds, some of which may be recycled to different heterocyclic species.

All the common monobasic (107) and dibasic esters (108) of tetrahydrofurfuryl alcohol have been prepared by conventional techniques; the dibasic esters and some of the mono esters are effective as primary or secondary plasticizers for vinyl polymers. Tetrahydrofurfuryl acrylate [2399-48-6] and methacrylate [2455-24-5], specialty monomers, have been produced by carbonylation (nickel carbonyl and acetylene) of the alcohol (109) as well as by direct esterification (110–112) and ester interchange (111).

Tetrahydrofurfuryl alcohol reacts with ammonia to give a variety of nitrogen containing compounds depending on the conditions employed. Over a barium hydroxide-promoted skeletal nickel–aluminum catalyst, 2-tetrahydrofurfurylamine [4795-29-3] is produced (113–115). With palladium on alumina catalyst in the vapor phase (250–300°C), pyridine [110-86-1] is the principal product (116–117); pyridine also is formed using Zn and Cr based catalysts (118,119). At low pressure and 200°C over a reduced nickel catalyst, piperidine is obtained in good yield (120,121).

Tetrahydrofurfuryl alcohol is oxidized to 2-tetrahydrofurfural [7681-84-7] in good yield by passage, with oxygen, over silver gauze at 500°C (123–124). With chromate oxidizing agents, lactones are also formed (124,125).

Tetrahydrofurfuryl alcohol is surprisingly resistant to hydrogenolysis; under vigorous conditions, however, cleavage of the ring or side chain occurs (126,127).

When tetrahydrofurfuryl alcohol is passed over a variety of alumina-containing catalysts at temperatures ranging from 200–500°C, it undergoes dehydration and ring expansion to 3,4-dihydro-2*H*-pyran, a major chemical intermediate for 1,5-difunctional open-chain aliphatic compounds (128–131) as well as a hydroxyl group protecting agent. The analogous 3,4-dihydro-2*H*-thiopyran is similarly prepared by reaction of tetrahydrofurfuryl alcohol with hydrogen sulfide over alumina at 300°C (132,133).

Manufacture. Tetrahydrofurfuryl alcohol is produced commercially by the vapor-phase catalytic hydrogenation of furfuryl alcohol. Liquid phase reduction is also possible.

Normal precautions for chemicals of mild toxicity are applicable to the safe handling and storage of commercial tetrahydrofurfuryl alcohol. Discoloration in storage rarely occurs if the proper precautions are observed; prevention of exposure to air will prevent autoxidation. The list price of tetrahydrofurfuryl alcohol (1997) is \$1.15 lb.

Uses. Tetrahydrofurfuryl alcohol is of interest in chemical and related industries where low toxicity and minimal environmental impact are important (134). For many years tetrahydrofurfuryl alcohol has been used as a specialty organic solvent. The fastest growing applications are in formulations for cleaners (135) and paint strippers (136), often as a replacement for chlorinated solvents (137). Other major applications include formulations for crop sprays, water-based paints, and the dyeing and finishing of textiles and leathers. Tetrahydrofurfuryl alcohol also finds application as an intermediate in pharmaceutical applications.

A major use of tetrahydrofurfuryl alcohol is as an ingredient in proprietary stripping formulations. These formulations are used broadly in the automotive industry to remove protective coatings, paint, and grime prior to final painting and finishing (138). Stripping formulations containing tetrahydrofurfuryl alcohol are also available for the lifting and removal of epoxy coatings (139). Tetrahydrofurfuryl alcohol is being widely used as a replacement for chlorinated solvents, especially in the electronics industry.

Because tetrahydrofurfuryl alcohol is virtually colorless, it is used in lacquer formulations for all colors as well as water-white clear products. More specifically, tetrahydrofurfuryl alcohol is a wetting dispersant for most pigments. It has a high boiling point, high toluene dilution ratio, and good miscibility with oils, eg, linseed and soya, and is an excellent solvent for a wide range of resins.

Tetrahydrofurfuryl alcohol is used as solvent and carrier for industrial and commercial cleaning formulations. Product applications include the commercial cleaning of aircraft, ships, and trucks as well as the industrial cleaning requirements for machine tools and dies. The solvent properties of tetrahydrofurfuryl alcohol and its low rate of evaporation are utilized in biocides and pesticides where it serves as solvent–carrier. In a recent paper, tetrahydrofurfuryl alcohol is compared to other solvents as an agricultural adjuvant (140). Tetrahydrofurfuryl alcohol has been cleared by the EPA for use in formulations sprayed on growing food crops. The fact that tetrahydrofurfuryl alcohol is totally miscible with water and easily biodegradable enhances its use in agricultural sprays used for pest control, weed control, and growth retardation.

Tetrahydrofurfuryl alcohol also has been cleared for use in California under Rule 66. It meets the rigid specifications for uv degradation in environmentally sensitive areas. In addition to domestic use, a substantial amount of tetrahydrofurfuryl alcohol is used as a solvent for agricultural chemicals in Europe.

Tetrahydrofurfuryl alcohol is a solvent and coupling agent for a phosphate-type insecticide used to control the gypsy moth. Esters of tetrahydrofurfuryl alcohol are used in preparations employed as insect repellents. Tetrahydrofurfuryl alcohol is also used as a solvent–carrier for an EPA-approved paper slimicide formulation. In this application, the exceptional solvent action of tetrahydrofurfuryl alcohol prevents separation of the

active ingredient, especially at low operating temperatures.

Tetrahydrofurfuryl acrylate and methacrylate reactive unsaturated monomers, are readily polymerized and easily cross-linked by exposure to heat, peroxide catalysts, or uv radiation.

Tetrahydrofurfuryl acrylate is used as a co-reactive viscosity depressant for vinyl terminated epoxy systems. It is also used in the formulation of uv-curable printing inks, coatings, paints, and adhesives. Tetrahydrofurfuryl methacrylate serves as a reactive curing agent in the peroxide-catalyzed production of nitrile rubber. The monomer functions as plasticizer during processing and polymerizes during cure to yield hard vulcanizates useful in electrical cable coatings.

Tetrahydrofurfuryl alcohol is used in elastomer production. As a solvent for the polymerization initiator, it finds application in the manufacture of chlorohydrin rubber. Additionally, tetrahydrofurfuryl alcohol is used as a catalyst solvent-activator and reactive diluent in epoxy formulations for a variety of applications. Where exceptional moisture resistance is needed, as for outdoor applications, furfuryl alcohol is used jointly with tetrahydrofurfuryl alcohol in epoxy adhesive formulations.

In dyeing applications, tetrahydrofurfuryl alcohol permits higher dye concentrations and wider latitude in solution temperature because of its high solvency for dyes. Through swelling action on synthetic fibers and polarity with respect to leather, tetrahydrofurfuryl alcohol facilitates deeper penetration of dye into the substrate circumventing mere surface tinting. As a dye leveling agent, tetrahydrofurfuryl alcohol aids in the attainment of evenness in color shade.

Other Furan Derivatives

Other furan compounds, best derived from furfural, are of interest although commercial volumes are considerably less than those of furfural, furfuryl alcohol, furan, or tetrahydrofurfuryl alcohol. Some of these compounds are still in developmental stages. Applications include solvents, resin intermediates, synthetic rubber modifiers, therapeutic uses, as well as general chemical intermediates.

Solvents. Compounds containing the furan ring are generally excellent solvents. Some are miscible with both water and with hexane. Presence of the ether oxygen adds polarity as well as the potential for hydrogen bonding. Ring-substituted derivatives of furfural and furfuryl alcohol are reactive solvents, similar to the parent compounds. Reaction of the aldehyde group on furfural or the alcohol group on furfuryl alcohol give products that are merely solvents, no longer reactive. Ethers of furfuryl alcohol are easy to prepare; most simple ethers have been synthesized. Ethylene oxide adducts [31692-86-1] are used in paint stripping formulations (136,141–143). The esters of furfuryl alcohol are effective solvents, but must be synthesized carefully owing to the reactivity of furfuryl alcohol with acids. Both esters and ethers of tetrahydrofurfuryl alcohol are excellent solvents and are easily prepared. Since double bonds are no longer present, these compounds are more stable than the corresponding furan derivatives. Tetrahydrofurfuryl alcohol–ethylene oxide adducts [31692-85-0] are also useful solvents for paint stripping formulations (136,141,143). 2-Methylfuran is a good solvent, but 2-methyltetrahydrofuran (METHF), is completely saturated and is less reactive, less toxic, and has greater potential for a number of solvent applications. It is a more convenient solvent than tetrahydrofuran for Grignard reagents (144), is higher boiling, and wet METHF is more easily recovered and made anhydrous for recycle and reuse. METHF has also found application as a solvent for other organometallic reagents (145), as well as in lithium batteries (146). Ultimately, when fully commercialized, the cost of METHF should approximate that of tetrahydrofuran.

Resins. As mentioned above, both furfural and furfuryl alcohol are widely used in resin applications. Another resin former, 2,5-furandimethanol [1883-75-6] (BHMF), is prepared from furfuryl alcohol by reaction with formaldehyde. It is usually not isolated because oligomerization occurs simultaneously with formation (competing reaction). Both the monomer and oligomers are very reactive owing to difunctionality, and are used primarily as binders for foundry sand (72) and fiberglass insulation (147,148).

Therapeutics. Compounds containing the furan or tetrahydrofuran ring are biologically active and are present in a number of pharmaceutical products. Furfurylamine [617-89-0] is an intermediate in the diuretic, furosemide. Tetrahydrofurfurylamine [4795-29-3] may also have pharmaceutical applications. 5-(Dimethylaminomethyl)furfuryl alcohol [15433-79-1] is an intermediate in the preparation of ranitidine, which is used for treating ulcers. 2-Acetylfuran [1192-62-7], prepared from acetic anhydride and furan is an intermediate in the synthesis of cefuroxime, a penicillin derivative. 2-Furoic acid is prepared by the oxidation of furfural. Both furoic acid [88-14-2] and furoyl chloride [527-69-5] are used as pharmaceutical intermediates. 2-Tetrahydrofuroic acid [16874-33-2] also finds application.

Rubber Modifiers. Derivatives of furan and tetrahydrofurfuryl alcohol are used in the polymerization of synthetic rubber to control stereoregularity and other properties (149,150).

Intermediates. 3,4-Dihydro-2H-pyran [110-87-2] is prepared by a ring-expanding dehydration of tetrahydrofurfuryl alcohol. It is used as a protecting agent for hydroxyl compounds and as an intermediate. 2-Methylfuran is a chemical intermediate for 5-methylfurfural [620-02-0] (151) and 3-acetyl-1-propanol [1071-73-4] (152), and is also used in perfume. α-Methylfurfuryl alcohol [4208-64-4] is prepared by the reaction of furfural with methylmagnesium bromide. It is an intermediate in the synthesis of maltol [118-71-8], which is used as a flavoring agent.

Health and Safety

As with all chemical compounds, the Material Safety Data Sheet (MSDS) for each of the specific furan derivatives should be reviewed before starting to work with these materials. Caution should be the keyword when handling chemicals. Additional information on toxic effects of most of these compounds can be found in RTECS (*Registry of Toxic Effects of Chemicals*), HSDB (*Hazardous Substances Data Bank* from the National Library of Medicine), and standard works on toxicology such as Patty's (153). Toxicology studies are taking place on a continuing basis with many chemicals; furan derivatives are no exception. New data may change the perspective on toxicity of these chemicals.

Over the years, many people have been exposed to low levels of these furan derivatives with no evidence of any significant long-term effects. Proper precautions, however, should be taken when working with these compounds as furfural, furfuryl alcohol, and furan are moderately toxic, tetrahydrofurfuryl alcohol is less so. Protective clothing should be worn; skin absorption is a likely route of entry. Eye protection is required. Proper ventilation is needed, especially with highly volatile furan. The odor threshold for these compounds is such that they can be detected at levels below the PEL (Permissible Exposure Level), and appropriate action taken as soon as the odor is detected. Specific recommendations from the manufacturer regarding exposure and protective measures should be followed. Since regulations change from time to time, up-to-date exposure limit recommendations from OSHA (Occupational Safety and Health Agency) or ACGIH (American Council of Governmental Industrial Hygienists) need to be consulted and followed.

Toxicity information is summarized in Table 3.

Table 3. Toxic Effects of Furfural and Main Derivatives^a

Property	Value for			
	Furfural	Furfuryl alcohol	Furan	Tetrahydrofurfuryl alcohol
oral toxicity	149 mg kg	451 mg kg	300 mg kg	2500 mg kg
inhalation toxicity	1037 ppm 1 h	692 ppm 1 h	3464 ppm	LC ₂ 3, 12650 ppm
dermal toxicity	620 mg kg	667 mg kg	not established	>5000 mg kg
skin irritation	slight	slight	not established	slight
eye irritation	moderate	moderate	not established	moderate
ACGIH carcinogen	not listed	not listed	not listed	not listed
IARC carcinogen	not listed	not listed	not listed	not listed
NIOSH carcinogen	not listed	not listed	not listed	not listed
NTP carcinogen	not listed	not listed	not listed	not listed
OSHA carcinogen	not listed	not listed	not listed	not listed

^a Ref. 154.

BIBLIOGRAPHY

"Furfural and Other Furan Compounds" in *ECT* 1st ed., Vol. 6, pp. 995–1008, by H. R. Duffey, The Quaker Oats Company, and H. J. Barrett, E. I. du Pont de Nemours & Co., Inc.; "Furfural and Other Furan Compounds" in *ECT* 2nd ed., Vol. 10, pp. 237–250, by A. P. Dunlop, The Quaker Oats Company; "Furan Derivatives" in *ECT* 3rd ed., Vol. 11, pp. 499–527, By W. J. McKillip and E. Sherman, The Quaker Oats Company.

1. A. P. Dunlop and F. N. Peters, *The Furans*, ACS Monograph 119, Reinhold Publishing Corp., New York, 1953.
2. M. J. Cook, A. R. Katritzky, and P. Linda, *Adv. Heterocycl. Chem.* **17**, 255 (1974).
3. C. W. Bird, *Tetrahedron* **48**, 335 (1992).
4. M. V. Sargent and T. M. Cresp "Furans" in D. Barton and W. D. Ollis, eds., *Comprehensive Organic Chemistry: The Synthesis and Reaction of Organic Compounds*, Vol. 4, Pergamon Press Ltd., Oxford, U.K., 1979, pp. 693–744.
5. A. Gandini "Furan Polymers" in J. I. Kroschwitz, ed., *Encyclopedia of Polymer Science and Technology*, 2nd ed., Vol. 7, John Wiley & Sons, Inc., New York, 1967, pp. 454–473.
6. A. Gandini, *Adv. Polym. Sci.* **25**, 47 (1977).
7. F. M. Dean, *Advances in Heterocyclic Chemistry*, **30**, 168; **31**, 237 (1982).
8. Dobereiner, *Ann.* **3**, 141 (1832); quoted in Ref. 1, p. 272.
9. Emmet, in *J. Pract. Chem.* **12**, 120 (1837); *Ann.* **28**, 249 (1939); quoted in Ref. 1, p. 272.
10. Stenhouse, *Phil. Mag.* **18**, 122 (1840); *Ann.* **35**, 301 (1840); quoted in Ref. 1, p. 272.
11. Baeyer, *Ber.* **10**, 355, 695, 1358 (1877); quoted in Ref. 1, p. 4.
12. Markwald, *Ber.* **20**, 2811 (1887); quoted in Ref. 1, p. 5.
13. Markwald, *Ber.* **21**, 1398 (1887); quoted in Ref. 1, p. 5.
14. Harties, *Ber.* **34**, 1488 (1901); quoted in Ref. 1, p. 5.
15. G. H. Mains, *Chem. Met. Eng.* **26**, 779 (1922).
16. R. G. Curtis and H. H. Hatt, *Aust. J. Sci. Res.* **A1**, 213 (1948).
17. A. P. Dunlop and F. N. Peters, Jr., *Ind. Eng. Chem.* **32**, 1639 (1940).
18. The Quaker Oats Company, Unpublished data.
19. T.-S. Li, S.-H. Li, J.-T. Li, and H.-Z. Li, *J. Chem. Res. Synop.*, 26–27, (1997); *Chem. Abst.* **126**, 144440 (1997).
20. V. V. Girdyuk, Yu. K. Kirilenko, L. A. Vol'f, and A. I. Meos, *Zh. Prikl. Khim.* **39**, 2601 (1966).
21. I. V. Kamenskii, S. M. Filimonova, Ngo Van Lam, B. Ya. Eryshev, and N. B. Krovyakova, *Plast. Massy* **5**, 5 (1973); *Chem. Abstr.* **79**, 79683n (1973).
22. Pais, C. G. Godwin; A. Keshavaraja; K. Saravanan, and P. Kumar, *J. Chem. Res. Synop.* **9**, 426 (1996).
23. N. Dekar, D. J. Kalita, R. Borah, and J. C. Sarma, *J. Org. Chem.* **62**, 1563–1564 (1997).
24. W. J. Chute and G. F. Wright, *J. Org. Chem.* **10**, 541 (1945).
25. H. Gilman and G. F. Wright, *J. Am. Chem. Soc.* **52**, 2550 (1930).
26. V. I. Sakhnenko, M. V. Sokolov, V. V. Kashmet, A. N. Chernobrovyi, *Khim.-Farm. Zh.* **27**, (3), 53–7 (1993) *Chem. Abst.* **121**, 160125 (1994).
27. H. Li and M. Gao, *Hebei Daxue Xuebao, Ziran Kexueban*, **12**, 78 (1992).
28. D. L. Williams and A. P. Dunlop, *Ind. Eng. Chem.* **40**, 239 (1948).
29. J. Z. Marcusson, *Z. Angew. Chem.* **32**, 113 (1919).
30. G. Illari, *Gazz. Chim. Ital.* **77**, 389 (1947).
31. Y. Nakamura and M. Saito, *Kogyo Kagaku Zasshi* **62**, 1173 (1959).
32. N. Galego, Ph.D. Thesis, University of Havana, 1975.
33. N. Galego and A. Gandini, *Rev. CENIC Cent. Nac. Invest. Cient. I Cienc. Fis.* **6**, 163 (1975).
34. N. Galego and F. Lopez, *Rev. Cubana Quim.* **6**, 107 (1992).
35. R. Sanchez, C. Hernandez and G. Keresztury, *Eur. Polym. J.* **30**, 43 (1994).
36. R. L. Whistler and E. L. Richards, "Hemicelluloses" in W. Pigman and D. Horton, eds., *The Carbohydrates*, Vol. IIA, Academic Press, Inc., New York, 1970, pp. 447–467.
37. G. O. Aspinall, *Polysaccharides*, Pergamon Press Ltd., Oxford, U.K., 1970, pp. 103–115.
38. *Making and Marketing Furfural; Added Value for Agro-Industrial Wastes*, International Trade Center UNCTAD GATT, Geneva, 1979.
39. J. E. Stone and M. J. Blundell, *Can. J. Res.* **28B**, 676 (1950).
40. R. Will, "Furfural," *Chemical Economics Handbook*, Stanford Research Institute, Menlo Park, Calif., Mar. 1994.
41. **Jpn. Pat. 78 19,693** (June 22, 1978), M. Futamoto, U. Kawabe, and S. Hosoki (to Hitachi Ltd.).
42. *AID (Air Information Division) Report 61–19*, Armed Services Technical Information Agency, NTIS (National Technical Information Service), Springfield, Va., Feb. 1961.
43. *AID Report 61-5*, Armed Services Technical Information Agency, NTIS, Springfield, Va., Jan. 1961.
44. D. A. Isacescu, I. Gavai, C. Stoicescu, C. Vass, and I. Petrus, *Rev. Roum. Chim.* **10**, 219 (1965); *Chem. Abst.* **63**, 16284 (1965).
45. A. A. Patel, Ashok, and S. R. Patel, *Eur. Polym. J.* **19**, 231 (1983); *Chem. Abst.*, **98**, 216122 (1983).
46. I. V. Kamenskii, N. V. Ungurean, B. M. Kovarskaya, and I. V. Itinskii, *Plast. Massy* **12**, 9 (1960); *Chem. Abst.* **55**, 15985 (1961).
47. **Hungarian Pat. 58780** (Mar. 30, 1992); *Chem. Abst.* **117**, 193648 (1993).
48. **U.S. Pat. 5,037,453** (Aug. 6, 1991), K. S. Narayanan and M. S. Ramakrishnan (to the Norton Co.).
49. **U.S. Patent 4,348,343** (Sep. 7, 1982), D. W. Akerberg, G. W. Huffman, and C. A. Rude (to The Quaker Oats Co.).
50. S. A. Suvorov, D. E. Denisov, V. F. Timofeev, G. I. Kuznetsov, A. A. Kortel, Yu. D. Kuznetsov, G. I. Borovikov, A. I. Shullman, and A. F. Radionov, *Ogneupory* **11**(11), 16 (1993).
51. A. P. Dunlop and F. N. Peters, Jr., *Ind. Eng. Chem.* **34**, 814 (1942).
52. V. Sunjic, J. Horvat, B. Klaic, and S. Horvat, *Kem. Ind.* **33**, 593 (1984); *Chem. Abst.* **102**, 151194 (1985).
53. J. Horvat, B. Klaic, Branimir, B. Metelko, S. Biserkaand, and V. Sunjic, *Croat. Chem. Acta*, **59**, 429 (1986).
54. A. Gandini, *ACS Symp. Ser. 433, Agric. Synth. Polym.: Biodegrad. Util.*, 195 (1990).
55. M. Choura, N. Belgacem, and A. Gandini, *Macromolecules* **29**, 3839 (1996).
56. R. H. Wojcik, *Ind. Eng. Chem.* **40**, 210 (1948).
57. **U.S. Pat. 2,754,304** (July 10, 1956), S. Swadesh (to The Quaker Oats Co.).
58. W. D. Wakefield, *Production* (Oct. 1976).
59. R. F. Frankenberg, *Foundry Met. Treat.*, 87 (Jan. 1979).
60. *Int. Mod. Foundry*, **2**, 2 (1978).
61. S. Pal, *Indian Foundry J.* **35**, 27 (1989); *Chem. Abst.* **112**, 143743 (1990).
62. **U.S. Pat. 4,495,316** (Jan. 22, 1985), D. R. Armbruster (to Acme Resin Corp.).

63. U.S. Pat. **4,543,373** (Sept. 24, 1985), R. W. Krawiec and J. E. Menting (to QO Chemicals).

64. U.S. Pat. **4,644,022** (Feb. 17, 1987) R. Iyer (to Acme Resin Corp.).

65. Eur. Pat. Appl. **698,432 A1** (Feb. 28, 1996) K. Kiuchi, S. Nakai, M. Sawa, M. Kato, M. Sakaiand, and S. Nomura (to Kao Corp.).

66. W. H. Buell, *Foundry* **89**(2), 64 (1961).

67. A. Dorfmueller, Jr., *Foundry* **90**(4), 54 (1962).

68. PCT Int. Appl. 96-36,448 (Nov. 21, 1996), K. K. Chang, T. E. Dando, and A. L. Haugse (to Ashland Inc.).

69. U.S. Pat. **5,607,986** (Mar. 4, 1997), Y. D. Kiom and A. L. Haugse (to Ashland Inc.).

70. U.S. Pat. **4,255,554** (Mar. 10, 1981), J. P. Wuskell (to The Quaker Oats Co.).

71. Eur. Pat. Appl. **653,260** (Jan. 11, 1995), J. W. Ward, R. A. Llitar, and B. E. Wise (to Borden Inc.).

72. R. H. Kottke and A. E. Bloomquist, *AFS Trans.* **86**, 215 (1978).

73. U.S. Pat. **4,543,374** (Sept. 24, 1997), J. E. Menting (to QO Chemicals Inc.).

74. U.S. Pat. **4,451,577** (May 29, 1984), W. W. Coss (to The Quaker Oats Co.).

75. M. B. Launikitis, *Manag. Corros. Plast. Series* **3**, 190 (1977).

76. R. H. Leitheiser, K. B. Bozer, and D. D. Watson, *Furan Resins for FRP Composites*, National Association of Corrosion Engineers, Annual Meeting, San Francisco, Calif., Mar. 14–18, 1977.

77. R. M. Webster, National Association of Corrosion Engineers, Annual Meeting, San Francisco, Calif., Mar. 14–18, 1977.

78. *Impregnating Formulations Based on Furan Chemicals*, Technical Bulletin No. 190, Chemicals Division, The Quaker Oats Company, Chicago, Ill., 1979.

79. Jap. Pat. Appl. **06171918 A2** (June 21, 1994), O. Hirai and K. Oota, (to Hitachi Chemical Co. Ltd.).

80. U.S. Pat. **4,619,706** (Oct. 28, 1986), D. G. Squires, L. Hundley, and R. A. Berry (to Texo Corp.).

81. U.S. Pat. **4,366,022** (Dec. 28, 1982), C. M. Carandang (to Amchem Products, Inc.).

82. U.S. Pat. **5,259,993** (Nov. 9, 1993), S. M. Short (to Cook Composites and Polymers).

83. Jap. Pat. Appl. **03062898** (Mar. 18, 1991), K. Kitazawa and M. Takeda (to Kao Corp.).

84. C. D. Hurd and A. R. Goldsby, *J. Am. Chem. Soc.* **54**, 2558 (1932).

85. Jpn. Pat. **58090548 A2** (May 30, 1983) (to Daicel Chemical Industries, Ltd.): *Chem. Abst.* **99**, 105121 (1983).

86. Jpn. Pat. **01301658 A2** (Dec. 5, 1989), T. Yamamoto and M. Iwasaki (to Asahi Chemical Industry Co., Ltd.); *Chem. Abst.* **112**, 198124.

87. Yu. K. Yur'ev, *J. Gen. Chem. USSR* **8**, 1934 (1938).

88. A. Norton, *Chem. Rev.* **31**, 319 (1942).

89. G. B. Butler, J. T. Badgett, and M. Sharabash, *J. Macromol. Sci. Chem.* **4**(1), 51 (1970).

90. B. Kamo, I. Morita, S. Horie, and S. Furusawa, *Polym. J.* **6**(2), 121 (1974).

91. N. G. Gaylord, S. Maiti, B. K. Patnaik, and A. Takahashi, *J. Macromol. Sci. Chem.* **6**, 1459 (1972).

92. N. G. Gaylord, M. Martan, and A. B. Deshpande, *J. Polym. Sci. Polym. Chem. Ed.* **16**, 1527 (1978).

93. N. A. Milas and W. L. Walsh, *J. Am. Chem. Soc.* **57**, 1389 (1935).

94. K. G. Akopyan, A. G. Dzhomardyan, A. P. Sayadyan, G. K. Garibyan, and N. M. Mordyan, *Arm. Khim. Zh.* **41**, 451–454 (1988); *Chem. Abst.* **110**, 212078 (1989).

95. Ger. Pat. DE **2710420** (Aug. 24, 1978), D. Degner, H. Nohe, and H. Hannebaum (to BASF).

96. R. E. W. Jansson and N. Tomov, *Chem. Ind. (London)*, 96–98 (1978).

97. Russian Pat. Ru **2,027,714** (Apr. 28, 1992), N. G. Zheltonog, G. E. Kramerova, N. N. Machalaba and V. E. Oleinikov, *Chem. Abst.* **124**, 86804 (1995).

98. N. N. Machalaka, G. F. Tereshchenko, A. G. Bazanov, and N. G. Zheltonog, *Gidroliz. Lesokhim. Prom-st*, 25 (1993).

99. K. J. Jung, Ph. Lejemble, Ph. Kalck, J. Molinier, and A. Gaset, *Inf. Chim.* **283**, 135 (1987).

100. K. J. Jung, A. Gaset, and J. Molinier, *Biomass*, **16**, 89 (1988).

101. U.S. Pat. **2,846,449** (Aug. 5, 1958), W. H. Banford and M. M. Manes (to E. I. du Pont de Nemours & Co., Inc.).

102. U.S. Pat. **3,021,342** (Feb. 13, 1962), D. G. Manly (to The Quaker Oats Co.).

103. R. Paul, *Compt. Rend.* **206**, 1028 (1938).

104. B. A. Keay and P. W. Dibble, in C. W. Bird, ed., *Comprehensive Heterocyclic Chemistry II*, Vol. 2, Elsevier, Oxford, U.K., 1996, p. 395.

105. Belg. Pat. BE **1,008,868** (Aug. 6, 1996), J. E. Shaw and W. E. Sattich (to Phillips Petroleum Co.).

106. U.S. Pat. **3,850,832** (Nov. 26, 1974), B. Wegemund, H. J. Lehmann and E. Schmadel (to Hankel & Cie GmbH).

107. R. T. Wragg, *J. Chem. Soc.*, 7162 (1965).

108. L. H. Brown and J. W. Hill, *J. Chem. Eng. Data* **5**, 56 (1960).

109. W. Reppe and co-workers, *Ann.* **582**, 1 (1953).

110. Jpn. Pat. JP**58174346**, (Oct. 13, 1983) (to Yokkaichi Chemical Co., Ltd.).

111. Jpn. Pat. JP**50050313** (May 6, 1975), K. Kimura, K. Suzuki, E. Kondo, N. Motoyama, G. Terada, and H. Mishina (to Toa Gosei Chemical Industry Co., Ltd.).

112. Jpn. Pat. JP**50013313** (Feb. 12, 1975), K. Suzuki, K. Kimura, Y. Kato, H. Mishina, and E. Kondo, (to Toa Gosei Chemical Industry Co., Ltd.).

113. U.S. Pat. **2,636,902** (Apr. 28, 1953), A. W. C. Taylor, P. Davies, and P. W. Reynolds (to Imperial Chemical Industries Ltd.).

114. Jpn. Pat. JP**60178877** (Sept. 12, 1985), S. Mori, T. Aoki, R. Hamana and Y. Nomura (to Mitsubishi Petrochemical Co., Ltd.).

115. L. S. Glebov, G. A. Kliger, A. N. Shuikin, and V. G. Zaikin, *Neftekhimiya* **36**, 344 (1996).

116. U.S. Pat. **3,238,214** (Mar. 1, 1966), D. G. Manly, J. P. O'Halloran, and F. J. Rice (to The Quaker Oats Co.).

117. J. H. Choi and W. Y. Lee, *Appl. Catal. A* **98**, 21 (1993).

118. Soviet Pat. SU**1524917** (Nov. 30, 1989), K. M. Akhmerov, D. Yusupov, U. D. Nazirova, and A. B. Kuchkarov, (to Tashkent Polytechnic Institute).

119. U. D. Nazirova and K. M. Akhmerov, *Dokl. Akad. Nauk. UzSSR* 40 (1988).

120. U.S. Pat. **3,163,652** (Dec. 29, 1964), D. G. Manly, J. P. O'Halloran, and F. J. Rice (to The Quaker Oats Co.).

121. Jpn. Pat. JP**60174776** (Sept. 9, 1985), S. Mori, T. Aoki, R. Hamana and Y. Nomura, (to Mitsubishi Petrochemical Co., Ltd.).

122. Brit. Pat. GB**593,617** (Oct. 21, 1947), J. G. M. Bremner and R. R. Coats (to Imperial Chemical Industries).

123. R. A. Karakhanov, V. Vagabov, N. P. Karzhavina, O. M. Ramazanov, Y. S. Mardashev, and M. M. Vartanyan, *Zh. Prikl. Khim. (Leningrad)* **54**, 454 (1981).

124. S. Baskaran, I. Islam, and S. Chandrasekaran, *J. Chem. Res. Synop.* 290 (1992).

125. S. Baskaran and S. Chandrasekaran, *Tetrahedron Lett.* **31**, 2775 (1990).

126. E. P. Goodings and C. L. Wilson, *J. Am. Chem. Soc.* **73**, 4801 (1951).

127. S. Landa and J. Mostecky, *Collect. Czech. Chem. Commun.* **20**, 430 (1955).

128. O. W. Cass, *Chemurgic Dig.* **8**(6), 6 (1949).

129. R. Paul, *Bull. Sec. Chim. France* **14**, 158 (1947).

130. O. W. Cass, *Ind. Eng. Chem.* **40**, 21C (1048).

131. U.S. Pat. **3,518,281** (June 30, 1970), D. C. Holtman (to Emery Industries, Inc.).

132. R. F. Naylor, *J. Chem. Soc.*, 2749 (1949).

133. Yu. K. Yur'ev and E. G. Vendel'shtein, *J. Cen. Chem. USSR* **22**, 751 (1952).

134. *QO Tetrahydrofurfuryl Alcohol*, Bulletin 206, QO Chemicals, part of Great Lakes Chemical Corporation, West Lafayette, Ind.

135. U.S. Pat. 5,514,294 (May 7, 1996), G. W. Bohnert and co-workers (to Allied Signal Inc.).
136. U.S. Pat. 4,737,195 (Apr. 12, 1988) C. M. Caranding and R. E. Koch (to Amchem Products, Inc.).
137. U.S. Pat. 5,128,057 (July 7, 1992), M. L. Bixenman and G. C. Wolf (to Kyzen Corp.).
138. U.S. Pat. 4,514,325 (Apr. 30, 1985), A. Russo and T. Mauthner (to J. Hall Co.).
139. U.S. Pat. 5,456,853 (Oct. 10, 1995), M. M. Myers (to Rust-Oleum Corp.).
140. K. J. Doyel and co-workers, C. L. Foy, ed., *Adjuvants Agrichemistry*, CRC, Boca Raton, Fla., 1992, pp. 225–234.
141. U.S. Pat. 5,259,993 (Nov. 9, 1993), S. M. Short (to Cook Composites and Polymers Co.).
142. U.S. Pat. 4,619,706 (Oct. 28, 1986), D. G. Squires, L. Hundley, and R. A. Barry (to Texo Corp.).
143. U.S. Pat. 4,666,002 (Dec. 28, 1981), C. M. Carandang (to Amchem Products, Inc.).
144. L. Poppe, K. Recseg and L. Novak, *Syn. Com.* **25**, 3993 (1995).
145. WO. Pat. 97/06097, (Feb. 20, 1997), T. Rathman and co-workers (to FMC Corp.).
146. S. Tobishima and co-workers *J. Appl. Electrochem.* **27**, 902 (1997).
147. U.S. Pat. 5,486,557 (Jan. 23, 1996), D. W. Akerberg (to QO Chemicals, Inc.).
148. U.S. Pat. 5,589,536 (Dec. 31, 1996), C. M. Golino, C. A. Rude and co-workers (to QO Chemicals, Inc., and Schuller International, Inc.).
149. U.S. Pat. 4,429,091 (Jan. 31, 1984), J. E. Hall (to Firestone Tire and Rubber Co.).
150. A. F. Halasa and W. L. Hsu, "New Ether Modifiers for Anionic Polymerization of Isoprene," *Polym. Prepr.* **37**(2), (1996).
151. Jap. Pat. 57091982 A2 (Nov. 28, 1980) (to Sumitomo Chemical Co., Ltd.).
152. V. N. Pavlychev and co-workers, *Khim. Prom-st.*, (7), 394 (1980).
153. G. D. Clayton and F. E. Clayton, eds., *Patty's Industrial Hygiene and Toxicity*, 4th ed., John Wiley & Sons, Inc., New York, 1991.
154. Great Lakes Chemical Corp., information summarized from Material Safety Data Sheets.

General References

Refs. 1, 7–9.
M. V. Sargent and T. M. Cresp "Furans" in D. Barton and W. D. Ollis, eds., *Comprehensive Organic Chemistry: The Synthesis and Reactions of Organic Compounds*, Vol. 4, Pergamon Press Ltd., Oxford, U.K., 1979, pp. 693–744.

R. H. Kottke
Great Lakes Chemical Corporation

HIGH PURITY GASES

High purity industrial gases are routinely delivered in large quantities having purities exceeding 99.999% (> 5nines pure). Other commodity materials, whether in liquid or solid form, much less typically have purity exceeding 99.9% . There are many applications for gases where purity even higher than 99.999% is required. To service these applications, the technology for manufacturing and delivering high purity gases has developed into a multibillion dollar business worldwide. Japan and the United States have the majority of the market.

There is no universally accepted definition of what purity levels correspond to high purity. However, gases having total impurities specified <1 ppm on a molar or volume basis must be manufactured and handled differently from regular gases if that specification is to be maintained. A good working definition of high purity is gases having certain individual impurities held to levels <0.1 ppm.

The technology for manufacturing and delivering high purity gases has largely developed to support the manufacture of advanced semiconductor materials needed as part of the overall process of manufacturing integrated circuits. Purity requirements in some of these high performance materials have reached the point where impurity concentrations in the starting gases of 1 part in 10¹² have been related to reduced yields and poor performance in the resulting integrated circuits. Depending on a volume, high purity gases can be delivered using either bulk systems, where a plant-wide distribution system is integrated with central gas storage facilities, or cylinders, where a short local distribution system is supplied from a single high pressure cylinder.

Gases used in the manufacture of semiconductor materials fall into three principal areas: the inert gases, used to shield the manufacturing processes and prevent impurities from entering; the source gases, used to supply the molecules and atoms that stay behind and contribute to the final product, and the reactive gases, used to modify the electronic materials without actually contributing atoms or molecules.

Nitrogen [7727-37-9], N₂; oxygen [7782-44-7], O₂; hydrogen [1333-74-0], H₂; argon [7440-37-1], Ar; and helium [7440-59-7], He, are classified as semiconductor bulk gases. The following gases are delivered in cylinders and are classified as semiconductor specialty gases: (1) silicon precursor gases such as silane [7803-62-5], SiH₄; dichlorosilane [4109-96-0], SiH₂Cl₂; trichlorosilane [10025-78-2], SiHCl₃; and tetrachlorosilane [10026-04-7], SiCl₄; (2) dopant gases such as arsine [7787-42-1], AsH₃; phosphine [7803-51-2], PH₃; diborane [19287-45-7], B₂H₆ ; and boron trifluoride [7637-07-2], BF₃; (3) etching gases such as nitrogen trifluoride [7783-54-2], NF₃; sulfur hexafluoride [2251-62-4], SF₆ ; boron trichloride [10294-34-5], BCl₃; hydrogen fluoride [7664-39-3], HF; chlorine [7782-50-5], Cl₂; chlorine trifluoride [7790-91-2], ClF₃ ; hydrogen bromide [10035-10-6], HBr; tetrafluorosilane [7783-61-1], SiF₄; and other halocarbons; and (4) reactant gases such as ammonia [7664-41-7], NH₃; hydrogen chloride [7647-01-0], HCl; nitrous oxide [10024-97-2], N₂O; tungsten hexafluoride [7783-82-6], WF₆ ; and carbon dioxide [124-38-9], CO₂.

Production and Purification

The separations processes used for manufacturing high purity gases are generally the same as those used for making lower purity products. Purification by distillation (qv) and adsorption (qv) are often used and, provided basic data are available, can generally be designed using standard engineering practices. Some separation processes that are not economical for the manufacture of lower purity gases do find applications in making high purity gases. Chemical conversion processes, where the impurities are converted into more easily separable forms through a selective chemical reaction, are often employed in point-of-use purifiers which typically process streams having less than five standard liters per minute flow rate.

Whereas the underlying separation or purification technology may be straightforward, the purity achieved is often far less than that which the separation processes are capable of producing. More often than not, recontamination by impurities released by the materials of construction used in the purification, storage, and delivery equipment represents the true limit to the purity that can be achieved in practice.

Generally, the purification methods used for manufacturing high purity gases can be divided according to volume throughput requirements and of course differ according to the chemical and physical properties of the gas being purified. For the manufacture of high purity gases on a large scale, the industry tends to employ a few well-known methods. When production or purification is required on a smaller scale, a greater number of separation methods and combinations of these methods are used in actual practice.

Bulk Gases. The bulk gases are usually characterized by high volume flow requirements in the manufacturing process. Historically these have consisted of nitrogen, oxygen, argon, hydrogen, and to a lesser extent, helium. Because these gases are manufactured on a large scale, small-scale supply schemes tend to be more of a distribution process for product already purified on a large scale. These small-scale distribution systems may, however, incorporate so-called onsite or point-of-use purifiers to ensure that any impurities picked up during distribution are removed.

Generally the bulk gases are also nonreactive with their containment systems. However, this is becoming less true because the need for large quantities of high purity HCl, trichlorosilane, and other corrosive gases has grown pushing these into the bulk gases category.

Nitrogen. Because of numerous applications in semiconductor manufacturing, high purity nitrogen is produced both at high volumes and at some of the lowest impurity levels seen for any of the high purity gases. Streams up to 10,000 nm³ h can be produced at impurity levels of 1 part in 10⁹. *Distillation.* All high purity nitrogen is manufactured from air using multistage cryogenic distillation (qv). Depending on the customer's flow rate and purity requirements, the original separation from air is carried out either locally on the customer's property or at a large air separation (qv) unit (ASU) located some distance away.

For volume throughput requirements larger than ~1000 nm³ h, self-contained and highly automated on-site generators are installed on the user's property. The gaseous nitrogen produced by the generator is then delivered throughout the user's plant using a dedicated high purity distribution system. Impurity contents on the order of 1 part in 10⁹ can be achieved in the delivered product using these systems.

For smaller volume requirements, nitrogen is manufactured and liquefied in a large merchant air separation unit owned by the gas supplier. Liquid nitrogen is transported by tanker truck and stored in a large cryogenic tank on the customer's site. When needed by the customer, liquid nitrogen is withdrawn from the tank, vaporized, filtered, and delivered throughout the plant using a dedicated distribution system. Impurity levels on the order of 1 part in 10⁸ can be achieved by trucking liquid nitrogen over distances exceeding 1500 km. However, product costs are determined largely by delivery distance and may be uneconomical at that range.

Chemical Conversion. In both on-site and merchant air separation plants, special provisions must be made to remove certain impurities. The main impurity of this type is carbon monoxide, CO, which is difficult to separate from nitrogen using distillation alone. The most common approach for CO removal is chemical conversion to CO₂ using an oxidation catalyst in the feed air to the air separation unit. The additional CO₂ which results, along with the CO₂ from the atmosphere, is then removed by a prepurification unit in the air separation unit.

At throughputs below 500 nm³ h, a wide variety of inert gas purification processes based on chemical conversion can be used to produce high purity nitrogen. For the most part, these processes employ chemical reactions to convert O₂, H₂O, CO, and CO₂ impurities into nonvolatile products. Typically, the impure nitrogen is passed through a bed of reagents, which may or may not require elevated temperatures to function, where the conversion reactions occur causing the impurities to remain behind in the bed of reagents. These processes require that the bulk of the oxygen in the feed be removed by some other method. For this reason, they are mainly used to convert standard purity grades of nitrogen into high purity ones.

For nitrogen flows above 10 nm³ h, it is important to minimize any heating or cooling of the process stream because of the energy costs. Heating for regeneration purposes does not carry this economic penalty. There are two types of chemical conversion processes in general use which meet this requirement. Both function effectively at room temperature and below and are capable of removing O₂, H₂O, CO, and CO₂ to levels below 1 part in 10⁹.

The first process utilizes a bed of nickel catalyst which has been regenerated with hydrogen to reduce the nickel content to metallic form. The finely divided metal then reacts with impurities and retains them in the bed, probably as nickel oxide in the case of oxygen or as physisorbed compounds for other impurities. Periodically, the bed is regenerated at elevated temperature using hydrogen to restore the metallic content. The nickel process can be used and regenerated indefinitely.

The second process utilizes a bed of high surface-area porous zirconium–iron alloy pellets. The bed is regenerated periodically at elevated temperature to allow impurities on the surface of the pellets to diffuse into the bulk. The renewed surfaces are effective for removing impurities from the gas stream until, with use, they once again become saturated. The zirconium–iron process has the advantage of not requiring hydrogen for regeneration; however, zirconium–iron pellets must be replaced once these have fully reacted.

Combined Distillation and Chemical Conversion. On-site generators using distillation are almost always combined with chemical conversion purifiers in large bulk high purity nitrogen supply systems. Practical issues, such as ensuring a continuous supply of inert nitrogen during an interruption of electrical power or when the generator is being serviced, dictate that liquid nitrogen be stored for use when the generator is unavailable. Often backup nitrogen must be obtained from standard commercial sources that are not high purity. Also, even if high purity liquid nitrogen from the generator can be accumulated for backup purposes, commercial purity liquid nitrogen may be the only source available if the generator is out of service for more than a day or so.

The delivery of high purity nitrogen from commercial purity liquid nitrogen sources is accomplished by using high capacity chemical conversion purifiers which are active only when the backup stream is in use. Even so, the equipment cost is high and every attempt is made to tightly integrate the chemical conversion purifier with the nitrogen generator.

Oxygen. High purity oxygen for use in semiconductor device manufacture is produced in relatively small quantities compared to nitrogen. There are two different purification processes in general use for manufacturing the gas: distillation and chemical conversion plus adsorption.

For applications in semiconductor device manufacture, it is important to remove the chemically nonreactive noble gas impurities such as krypton, xenon, and particularly radon. Through spontaneous decay, radioactive isotopes of these gases in the purified oxygen can cause a wide range of radioactive daughter products to be incorporated into the electronic device materials. Further decay of the radioactive daughter elements provides a continuous source of ionizing radiation inside the device that can lead to a type of defect known as soft errors, particularly in large dynamic memories.

Distillation. As for nitrogen, all high purity oxygen is derived from air through the air separation (qv) process using cryogenic distillation (qv). Generally, air separation units that manufacture commercial purity oxygen also remove nitrogen and other light impurities to levels low enough for high purity applications. Argon is still present at concentrations above 1 part in 10⁵, but is not considered an impurity for most applications of high purity oxygen. Heavy impurities, typically hydrocarbons and noble gases, are usually not removed from commercial purity oxygen. To make high purity grades, an additional distillation column is added to the process. Some on-site nitrogen generators are also able to generate a small stream of high purity oxygen.

Chemical Conversion and Adsorption. Where additional distillation is not practical, hydrocarbons and heavy noble gases can also be removed by combining chemical conversion with adsorption. Commercial purity oxygen is passed through a high temperature bed of oxidation catalyst; usually combinations of platinum and rhodium on a high surface area alumina support. Hydrocarbon impurities are oxidized to CO₂ and H₂O. Temperatures in excess of 700°C are needed in the catalyst bed to ensure complete oxidization of methane. Because the process flow rates are typically small, these temperatures can be maintained without excessive energy costs. The CO₂ and H₂O products from catalytic oxidation of hydrocarbon impurities are removed using a temperature swing adsorption (TSA) process. The adsorbent is typically one of the molecular sieves (qv), such as 13X. If designed appropriately, the adsorption process can also be used to remove the heavy noble gases. Usually, however, only radon removal is necessary. Because radon has a 3.8-d half-life, it is only necessary to slow its transit through the adsorber bed for sufficient time for natural decay to reduce the concentration to an acceptable level.

Argon. High purity argon has many applications as an inert gas during the manufacture of semiconductor devices. High consumption rates are typical during processes which deposit high purity silicon epitaxial films onto polished wafers. Lower volume applications range from silicon crystal pulling to sputter deposition of metal films. In all these applications, nitrogen is a reactive impurity which, in addition to O₂, H₂O, CO₂, CO, and all hydrocarbons, must be removed from the argon to low levels.

Distillation. Conventional purity argon is separated from air using a combination of distillation and chemical conversion. High purity argon is made the same way. However, more stages of distillation are designed into the low pressure and pure argon columns in the air separation plant. The improvement in the efficiency of the low pressure column typically reduces the nitrogen impurity to below 1 ppm. Because argon is found at a concentration of only 1 vol % in air, it is not practical to supply argon from on-site generators.

Chemical Conversion. Except for control of nitrogen impurity levels, the same chemical conversion methods used for nitrogen purification at low flow rates can also be used for argon purification. Although used less commonly for argon purification than for nitrogen purification, these chemical conversion methods are applied in point-of-use purifiers located close to where the gas is consumed.

There are also chemical conversion purifiers which remove nitrogen in addition to other more reactive impurities. These purifiers require elevated temperatures to function and consequently are restricted to small process flow rates of typically a few liters per minute.

Hydrogen and Helium. Whereas hydrogen and helium are very different chemically, these gases have low boiling points and are normally liquefied during manufacture. Because the boiling points are so low, even very small amounts of trace impurities tend to freeze and form solid deposits. To prevent formation of these deposits, which can eventually plug process lines, trace impurities must be removed prior to liquefaction. Similar methods are used to purify hydrogen and helium prior to liquefaction.

Hydrogen. High purity hydrogen is usually delivered and stored as a cryogenic liquid and vaporized when needed. This vaporized liquid seldom needs any further processing to meet high purity specifications. However, in situations where liquid hydrogen cannot be used, high pressure hydrogen gas can also be delivered and further purified on-site to meet high purity specifications.

On-site purification of gaseous hydrogen is accomplished using combinations of chemical conversion, cryogenic adsorption, and palladium membrane processes depending on the purity and throughput requirements (see MEMBRANE TECHNOLOGY). For small hydrogen streams, stand-alone hydrogen purifiers based on selective permeation through palladium membranes are available. These units take advantage of the high permeability of hydrogen through palladium metal. Generally, the membranes are heated to get higher processing capacity per unit area. However, provisions must then be made to remove hydrogen from the membrane prior to cool-down to prevent formation of small cracks which then allow impurities to leak through the membrane. Small to large hydrogen streams can be purified by adsorption carried out at cryogenic temperatures, usually near the boiling point of liquid nitrogen. Activated charcoal and silica gel are both effective adsorbents for purification. However, if any significant quantity of oxygen must be removed, activated charcoal should not be used because of the potential for an explosive chemical reaction with the adsorbed oxygen during regeneration. Regeneration of the adsorbent is accomplished through a process of heating and reverse purging. Large hydrogen streams can be purified using a room-temperature palladium catalyst bed to convert oxygen impurity into water followed by cryogenic adsorption. Because the oxygen impurity has been removed, the higher adsorption capacity of activated charcoal can then be utilized without danger of explosion.

Helium. High purity helium is usually not required in large quantities and is therefore not commonly delivered as a cryogenic liquid. Instead, high pressure cylinders are filled from a liquid helium source by the gas supplier and then transported to the customer. When high purity helium is required, the high pressure gaseous helium is processed through an on-site purifier.

Because helium is not chemically reactive, the same chemical conversion processes used for purification of nitrogen and argon are also applied to helium purification. Additionally, cryogenic adsorption processes are used to remove the nonreactive impurities. As for hydrogen, the same precautions must be taken against possible adsorbent explosions when dealing with significant amounts of oxygen impurity.

Helium is separated from helium-bearing natural gases usually, but not always, in the process of removing nitrogen to improve the fuel value of the

natural gas. Thus, in a sense, the principal part of commercial helium is as a by-product. Natural gas components include methane and heavier hydrocarbons, helium, nitrogen, water vapor, carbon dioxide, sometimes hydrogen sulfide, small amounts of argon, and traces of neon and hydrogen. Water, carbon dioxide, and sulfides are first removed by scrubbing with monoethanolamine and diethylene glycol, followed by drying with alumina. Then the natural gas is concentrated in helium as the higher boiling hydrocarbons are liquefied and collected. Crude helium, concentrated to about 70% and containing other permanent gases, undergoes a cryogenic purification where the majority of nitrogen and argon are liquefied. Activated charcoal operating at liquid nitrogen temperatures or below is used to adsorb all remaining nonhelium gases. Helium that exceeds 99.9999% in purity results.

Specialty Gases. The specialty gases are generally more reactive than the bulk gases and usually have low volume flow requirements in their applications. Historically, these have been delivered almost exclusively in standard compressed gas cylinders. However, as the need for increased quantities has arisen, bulk specialty gas supply systems utilizing larger ton-sized containers and tube trailers are being developed.

Purification of specialty gases can be divided into two areas: purification done by the gas supplier on a bulk scale prior to filling the cylinder or other delivery container, and purification carried out by the consumer on a point-of-use scale generally just prior to use.

Bulk Purification. Many specialty gases originate as by-products or low purity intermediate chemicals produced during the course of manufacturing something else. For example, most of the high purity silane produced in the United States originates as an intermediate in the manufacture of metallurgical silicon.

In some cases the chemical manufacturer purifies a portion of this intermediate stream to make a high purity product. In other cases, the chemical manufacturer sells a low purity product to a gas company and the gas company purifies it to make a high purity product. In both cases, purification is done on a continuous basis, rather than cylinder by cylinder. The purification processes tend to utilize standard methods.

Distillation. Processes which utilize either simple liquid vapor flash processes or multistage distillation tend to be used often for purification of bulk specialty gases. The fact that the separation process is well understood goes a long way toward reducing overall risk when the gases are difficult to handle. Distillation has been applied to the manufacture of flammable gases, for example ultrapure silane, as well as many corrosive gases.

Adsorption and Chemical Conversion. In some cases, removal of moisture or oxygen added by small amounts of air contamination is all that is necessary to make a gas suitable for high purity applications. For example, high purity HCl is available in tube trailer quantities but may require removal of moisture inadvertently added during transport before it is distributed to process equipment within a semiconductor plant. With a limited objective, it is usually most effective to use an adsorption process. As long as the total mass of impurity which must be removed is low enough, the process may be designed either with or without the capability for repeated regeneration.

Chemical conversion processes can also be used for moisture and oxygen removal. These tend to be the same ones developed for the smaller point-of-use purifiers. Consequently there is little economy of scale and they are seldom able to be regenerated.

Point-of-Use Purification. For the user of cylinder quantities of reactive specialty gases, there are only a limited number of ways to remove impurities and obtain high purity. Specialized point-of-use purifiers have been developed that purify small streams of many important reactive gases. Whereas these point-of-use purifiers cannot remove all important impurities, they are usually effective for removing the contamination added by the users' gas distribution system, mostly air and moisture.

One important class of point-of-use processes utilizes a porous polymer containing reactive metals. Variations in the metal and polymer chemistry are made to optimize the process for different gas applications. This is an active area of development and purifiers are available for most of the principal specialty gases.

Delivery and Control

Once a gas has been purified, it must be brought to its intended point of use without being degraded by the addition of excessive contamination. This transport or delivery process can be technically challenging. Often, contamination is unintentionally added even before the gas has left the purifier during its original manufacturing process. This is especially true for the corrosive specialty gases, where minimizing contamination by iron and other metallic species arising from the purification system is a primary concern.

Technologies to prevent recontamination during the delivery process have been an intense area of development. Work in the 1970s focused mainly on preventing contamination of high purity inert gases such as N₂ and Ar by atmospheric leaks and the release of atmospheric impurities which had been adsorbed on the inner surfaces of delivery systems during assembly. More recently, work has focused on preventing contamination of corrosive gases, such as HCl and HBr, with contaminants that contribute metallic impurities such as Fe and Ti to the electronic materials.

Delivery methods for high purity gases can be divided both according to chemical reactivity with respect to the containment system and, to a less significant degree, according to volume throughput requirements. Many highly flammable gases such as H₂ and SiH₄ are still inert with respect to the containment systems. Even though special provisions must be made because of flammability, the technology used to deliver flammable gases is similar to that employed for inert gases.

Bulk Gases. Attaining high purity gases where they are used requires a suitable gas distribution system. To achieve a high purity distribution system, there must be an absence of dead zones, external leakage, outgassing, and particulate contamination. Dead zones are obvious sources of carry-over contamination and must be designed out of the system. External leakage can arise from the diffusion of air through small holes, which takes place despite a reverse pressure gradient. Leakage can also occur because of permeation, especially through organic polymers used for gasketed seals. *Outgassing* is a general term applied to the release of any gaseous contaminants from the materials of construction. It can result from impurities adsorbed on surfaces and dissolved within the bulk material of both system components and particulate contamination.

Over the years, substantial technology has been developed for minimizing the recontamination of high purity gases by the distribution system. The use of high cleanliness components having low outgassing surfaces is extremely important. Low permeability materials, typically 316L stainless steel, must be employed throughout the system. Minimum dead volume designs and leak-tight fabrication techniques are essential. High efficiency filtration (qv), using low permeability materials to reduce outgassing from adsorbed dissolved contaminants, is employed frequently. A typical bulk gas tanker features electropolished 316L stainless steel inner vessels and piping, as well as proprietary tank fill lines, bellows or diaphragm-sealed valves, and cryogenic filters in the product delivery line. Such a vessel and piping system minimizes welds, oxidized joints, and rough surfaces that might trap or generate particulate contamination.

Specialty Gases. The purity of specialty gases depends on the systems and procedures adopted by the distributors for bulk gas supply and cylinder preparation, filling, and delivery. The distributors need to follow strict cylinder selection, preparation, and filling procedures to ensure the quality of the products. Most of the precautions taken into consideration in the bulk gases delivery system are also applied for specialty gases to eliminate recontamination.

Analysis and Certification

Ensuring that the purification and delivery processes are working properly is essential to successful applications of high purity gases. Gas analysis technology is especially critical during startup of high purity processes. Once a process is demonstrated to be operating properly, it can generally be maintained by following rigorous operational protocols and using statistical process control techniques to identify deviations from these protocols.

Particulates. Particulate impurity and particulate control are significant issues in virtually all applications of high purity gases. In the semiconductor industry, gas specifications requiring <350 particles m⁻³ (10 particles ft⁻³) greater than 0.1 μm in diameter are typical. Separation of particulate impurities is an important part of the process for manufacturing high purity gases and the most common approach is through filtration (qv) technology. As a separation process, filtration is understood well enough that off-the-shelf filters can remove particles down to the limits of detection available in analytical instruments.

Description of Analytical Methods. Procedures for analyzing both gas and particulate impurities in high purity bulk gas products such as nitrogen, argon, oxygen, helium, and hydrogen are available. Typical gaseous impurities include oxygen, moisture, carbon monoxide, carbon dioxide, total hydrocarbons (THC), argon, and nitrogen analyzed to the low ppb level. Particle impurities are analyzed to the 0.1-μm level.

Recommended Equipment. A list of analyzers (stand-alone or in a mobile cart) and level of detection for specific gas and particle analysis follows:

Instrument	Detection level	Detection mechanism
atmospheric pressure ionization mass spectrometer (apims)	<1 ppb	
moisture analyzer	<5 ppb	dew point or microbalance
oxygen analyzer	<1 ppb	electrolytic
H ₂ , CO, CO ₂ , THC analyzer	<10 ppb	gc-RGA
		gc-FID with methanizer
Ar, N ₂ analyzer	<50 ppb	gc-DID, gc-HID
particle counter	>0.1 μm	LPC

The operation manual accompanying each analyzer should be consulted for the recommended startup, installation, and calibration procedures. *Gaseous Impurities.* Instrument calibration (excluding the apims) typically consists of two steps: zeroing and spanning. Zeroing is accomplished by allowing the analyzer to sample a gas, the contaminant level of which is below the lower detection limit of the analyzer. This is called a zero gas. A typical method for generating zero gas is to take the actual sample gas and run it through a gas purifier.

In some analyzers, background readings associated with noise or minute internal levels of impurities can be removed electronically. Other analyzers may not allow electronic zeroing. Background levels must still be checked using the zero gas. For gas chromatographs (gc), it is important to determine that there are no background peaks that can interfere with the integration of true impurity peaks. If large background levels are evident, the system must be checked for leaks and or more purging time must be allowed.

Spanning, accomplished using a sample gas containing a known volume concentration of impurity, is performed at levels that are the same order of magnitude as the required detection. The actual span concentration is selected so that the majority of expected measurements fall at or within its value.

As was the case for the zeroing procedure, some analyzers may need to be electronically spanned. The analyzer is allowed to sample the span gas until a stable reading is obtained and then this span concentration is entered. This particular method represents a single-point calibration. Some analyzers require that more than one span concentration be sampled so that a calibration curve can be generated, known as a multipoint calibration. For gas chromatographs the peak area counts associated with known concentrations are used to generate the calibration curve. For low ppb detection, accurate span gases can only be produced using a high purity blending system. This procedure involves taking a cylinder containing between 1 and 5 ppm volume concentration of impurity and diluting to ppb concentrations with zero gas. For best results, the cylinder matrix gas and zero gas must be of the same species as the gas to be analyzed. It is good practice to recheck the analyzer's zero after spanning is completed.

Single-point calibrations and zero determinations are performed whenever instruments are shut down and or relocated to a different sampling location. Should the recalibration vary by more than 10% of the value obtained during the initial calibration, the analyzer is recalibrated until stable calibration data are obtained.

The calibration procedures for an atmospheric pressure ionization mass spectrometer (apims) involves the generation of separate calibration curves for each of the monitored impurities. Because of the instrument's high level of detection sensitivity (ppb to <ppb), a dedicated, high purity blending system must be utilized to minimize background and produce accurate low ppb calibration gases. To carry out calibration, apims response is measured for a zero gas and at least five blended impurity concentrations. These concentrations are evenly distributed in the range of 0–50 ppb. The ion-current response recorded at each concentration for the individual impurity is normalized against the total ion-current response. The data are expressed as a percentage in terms of relative ion intensities, $R_a = 100 I_a / I_T$, where R_a is the relative ion intensity of impurity a ; I_a the ion-current response of impurity a ; and I_T the total ion-current response. The relative intensity response data is regressed against blended impurity concentration data to produce the linear model, $R_a = k_a (C_{am} + C_0)$, where C_{am} is the blended impurity concentration of impurity a ; C_0 , the background impurity level; and k_a the multiplication constant. Possible sources of background response include instrument noise, sample system outgassing, or interference from other impurity response signals. Proper setup, purging, and operation of the instrument should reduce background levels well below 1 ppb.

Rearranging the above expression yields impurity concentration as a function of relative intensity, $C_a = (R_a / k_a) - C_0$, where C_a represents both sample concentration and any background effects. The stability of the calibration must be confirmed at least every two weeks by analysis of a known mixed impurity standard.

Particulate Impurities. Particle counters require factory calibration every two years. In addition, the background signal associated with both the instrument and its sampling system (tubing, sampler, sample valves, regulators, and fittings) must be quantified so that it may be subtracted from sample measurements. To accomplish this, an absolute filter (<0.01 μm rating) is employed. The absolute filter removes all particles entering the sampling system, so any particle registered by the counter can be directly attributed to the sampling system or instrument noise.

To measure background, the absolute filter is placed at the beginning of the sampling system and purged for a minimum of one hour at a rate of 100 standard liters per minute (SLPM). After the purge, the sampling system is plumbed up to the particle counter and a minimum of 0.14 standard cubic meters (5 SCF) of gas is analyzed in 0.003 or 0.0003 standard cubic meter (0.1 or 0.01 SCF) increments. The average of these runs is accepted as the system background. It is recommended that the background be checked again at the completion of certification measurements for a particular system. The background varies depending on the matrix being analyzed, therefore when switching matrices, the background must again be quantified.

Sampling and Analysis Guidelines. As a general safety consideration, all gases should be vented to an external area and whenever possible, inert gases should be used as the test gas for piping systems.

Sample Line Connection Procedure. The following procedure is used for connecting sample lines to sampling ports. For best results, a one-eighth in. (0.318 cm) diameter 316SS electropolished tubing having one-fourth in. (0.635 cm) end fittings is recommended for sampling gas for both gaseous and particulate impurity analyses. The one-eighth in. diameter tubing allows a larger degree of flexibility than does the one-fourth in. tubing. Butt weld glands that adapt one-fourth in. end fittings to one-eighth in. tubing are available (Swagelok). The actual orbital welding can be performed by fabricators.

The following are mandatory: (1) a new pair of clean, dry, plastic, disposable gloves must be worn during this procedure to prevent the potential contamination of the gas system by oils, dirt, and moisture present on the hands; (2) prior to making a connection, all required materials must be gathered and close at hand, including all tools, gloves, gaskets, adapter fittings, sample line, etc; (3) all sampling components must be precleaned prior to assembly, and these components must be kept clean at all times with ends capped and stored in cleanroom plastic bags; (4) appropriate precautions must be taken if making a connection to a hydrogen system, which includes grounding all components prior to and after making the connection and confirming that there is no ignition source or oxidizer near by (area must be isolated with red "Hazard Area" tapes); (5) sample connections are made to valved ports only; (6) for bulk gases other than hydrogen, the valve is opened slightly to ensure that a flow comes out the port once its cap is cracked open (for hydrogen, it is recommended that the connection be made with the valve closed and the next step omitted); (7) standing to the side of the port (ie, not in front of the line of flow), the analyst should crack open the fitting slightly and ensure that gas is flowing out of the fitting, use the valve to adjust the flow leaking past the fitting to a substantial but manageable level, and if there is no flow with the valve opened, the fitting must be resealed immediately and the cause determined; (8) the cap and old gasket are removed and the sample line with new gasket connected; and (9) prior to connection to the analyzer distribution manifold, the sample line must be purged. The purge flow rate should be as high as possible, given all safety and gas consumption constraints. Sample line purging should be performed for at least 30 minutes. If the connection is made to a hydrogen system, this purging step is eliminated and the sample tubing immediately isolated until such time as it can be connected to the analyzers.

Sample Analysis. The following procedure is used for sample analysis: (1) after sample connections are made and the sample tubing purged for 30 minutes, the tubing is connected to the sampling manifold which serves to distribute sample gas to each of the analyzers, and should analyzers have specific pressure

and or flow requirements, a means of regulating these parameters should be incorporated into the manifold (suggested design: high purity regulator attached to a multivalve manifold having flow meters on each flow-critical leg); (2) the sampling system and analyzers are then purged for at least two hours (if this is the first use of the system, or if it has been shut down for a significant time, longer purge times will likely be required to reach representative level). It is crucial that the analytical system be continuously purged or effectively isolated when moving to a new point to avoid contamination and a mobile analytical cart with a purified, on-board purge is ideal for maintaining high purity conditions; (3) when analyzing multiple locations, it is recommended that separate sample lines be set up purging in advance to reduce analysis cycle time, but this should only be done when all safety considerations and gas consumption factors are taken into account; (4) repeated analyses are then performed for approximately one hour to determine the stability of the analyzers and if the required specifications have been met; (5) it is recommended that five consecutive analyses showing the impurities within specification be completed (average concentrations are recorded on a standard certification form given the following information: customer facility, sampling location, analyst, time and date, impurity specification, measured impurity concentration for each completed analysis, measured impurity concentration mean, and analyzers employed); (6) if any impurity fails to meet specification, the analyst should check for leaks, check the calibration, and or increase purge flow (slowly) if safe to do so. Another option is to cycle-purge the system by repeatedly closing the vent valve to pressurize and then opening the valve wide to rapidly depressurize. The analyzers must be isolated during this process and care should be taken not to allow the system pressure to fall so low (eg, <69 kPa (<10 psig)) that air can gain access and contaminate the system; (7) after the above six steps have been taken, sampling conditions are returned to normal and the system retested; (8) gas chromatographic analyses are based on a direct comparison of the area or peak heights of the individual impurity against its respective standard calibration curve, however, if the concentration of impurity found in the sample is less than the lowest standard evaluated, then an extrapolated linear regression value may be determined; (9) the particle count for a measured gas is the difference between the count taken during the analysis of the sample and that measured for the sampling system background according to standard statistical procedures; and (10) particle measurements are generally accomplished more effectively when performed separately from vapor impurities because a higher level of purging is usually required for particle analysis and a complex sampling system can generate additional particles under these conditions.

Applications of High Purity Gases

The applications of high purity gases are primary in the semiconductor industries. From 1991 to 1995, the North American semiconductor bulk gas sales increased from U.S. \$214 to \$252 million (annual growth rate of 4.2%) and specialty gas sales increased from U.S. \$78 million to \$169 million (annual growth rate of 21.4%).

In addition to the microelectronics industry, other applications for high purity oxygen include fiber optics manufacturing, production of pharmaceuticals, and usage as calibration media in research and development laboratories, and in the pollution control field. Applications for high purity hydrogen include oxidation processing and epitaxial growth for both silicon and gallium arsenide.

Argon has become an important commodity owing to its ability to remain inert even under extreme conditions of temperature and pressure. High purity argon is likely to be used in the high technology fields of electronics, fiber optics, research and development, powder metal spraying, and hot isostatic pressing.

Other applications of high purity specialty gases include hydrogen bromide for etching single-crystal silicon, polysilicon, and aluminum; nitrogen trifluoride as a fluorine source for *in situ* cleaning processes for chemical deposition equipment, semiconductor etching and deposition, and high energy chemical lasers; and sulfur hexafluoride as a key etching material in certain semiconductor manufacturing processes because it has a selectively high etch rate of silicon compared to silicon dioxide or silicon nitride.

The primary driving force for high purity gases has been the increasing purity demands from high technology industries such as ceramics, optical fibers, and silicon wafer fabrication. In 1985, the typical impurity requirement in high purity nitrogen for wafer fabrication was on the order of 500 to 100 ppb on a molar or volume basis. The impurity requirement changed to the range of 10 to 1 ppb in 1990 and to 1 to 0.1 ppb in 1995. Future demands for even higher purity gases are expected.

In addition to purification to remove contaminants, accurate analyses of extreme purity gases and selection of storage and delivery equipment to minimize recontamination, all the while keeping costs down, pose serious challenges for the manufacturers of high purity gases. One analytical breakthrough in the 1990s was the development of atmospheric pressure ionization mass spectrometers that monitor <ppb levels of impurities continuously. Further improvement of analytical techniques is a focus of industrial research and development facilities.

BIBLIOGRAPHY

General References

R. DiNapoli and A. M. Sass, in J. J. McKetta, ed., *Encyclopedia of Chemical Processing and Design*, Vol. 31, Marcel Dekker, Inc., New York, 1990, p. 236.
W. H. Whitlock, "The Ultra-High Purity Challenge", in *Separation of Gases, Proceeding of the Fifth BOC Priestley Conference*, Birmingham, U.K., Sept. 19–21, 1989, Royal Society of Chemistry, 1990.
M. Finnergan, *CryoGas Int.*, 13 (Apr. 1996).
H. A. Grieco, *CryoGas Int.*, 20 (Apr. 1996).
J. D. Borkman, W. R. Couch, and M. L. Malczewski, *Microcontamination*, 23 (Mar. 1992).
S. D. Cheung, G. L. Mooney, and D. L. Jenson, *Micro*, 59 (Oct. 1995).
N. M. Chodhury and L. J. Mostowy, *Micro*, 33 (Sept. 1995).
J. M. Davidson and T. P. Rhane, *Microcontamination*, 34 (Mar. 1987).
E. F. Ezell, A. Athalye, and M. Chigrinskiy, *Microcontamination*, 93 (July Aug. 1996).
J. Hart and A. Paterson, *Microcontamination*, 63 (July 1994).
G. Kasper, H. Y. Wen, and H. C. Wang, *Microcontamination*, 18 (Jan. 1989).
K. Siefering, H. Berger, and W. Whitlock, *Microcontamination*, 31 (Sept. 1992).
Ibid., p. 23 (Nov. 1992).
K. Siefering, W. Whitlock, and A. Athalye, *Microcontamination*, 41 (Mar. 1994).
R. M. Thorogood, A. Schwarz, W. T. McDermott, and C. D. Holcomb, *Microcontamination*, 28 (Aug. 1986).
P. Burggraaf, *Semicon. Int.*, 70 (Mar. 1990).
R. Iscoff, *Semicon. Int.*, 52 (Mar. 1991).
Ibid., p. 42 (Apr. 1992).

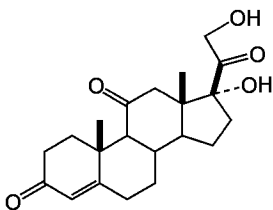
Walter H. Whitlock
Edward F. Ezell
Shuen-Cheng Hwang
The BOC Group, Inc.

HORMONES, ADRENAL-CORTICAL

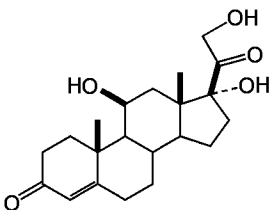
Since the introduction of cortisone (1) (1948) and hydrocortisone (2) (1951), adrenal-cortical hormones have remained an important and unreplaced drug class. Though not without adverse effects, these compounds have continued to be the drug of choice in the treatment of afflictions ranging from the moderate skin rash to severe acute inflammatory disorders, and are included in many other therapeutic regimes.

The adrenal cortex releases both mineralocorticoids (from the zona glomerulosa) and glucocorticoids (from the zona fasciculata reticularis). In addition, some androgenic and estrogenic steroids are synthesized by the adrenal gland. Once released by the adrenal cortex, the primary endogenous function of glucocorticosteroids is in influencing carbohydrate and protein metabolism. Mineralocorticoids regulate sodium reabsorption in the collecting tubules of the kidney. As with the mineralocorticoids, the generation of glucocorticosteroids is intricately balanced; where the production and regulatory process malfunctions, the result is either an excess (eg, Cushing's syndrome) or a deficiency (eg, Addison's disease) in glucocorticoid levels.

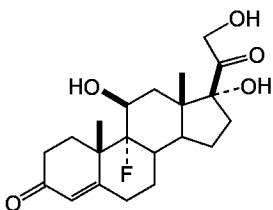
However, corticosteroids and their metabolites (1) were early recognized as possessing powerful antiinflammatory and immunomodulatory properties. Even prior to 1950, reports of the antiarthritic properties of cortisone (1) by Hench and co-workers (2) indicated the potential for these compounds to reduce the suffering of patients with inflammatory diseases. This awareness, combined with the first synthesis of naturally occurring glucocorticoids (11-desoxycorticosterone), led not only to the massive increase in research in the area of steroid synthesis and physiology, but to a Nobel prize in 1950 for early steroid pioneers Hench, Reichstein, and Kendall.



(1)



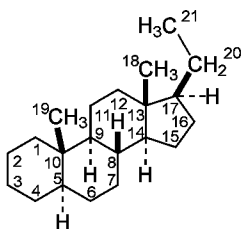
(2)



(3)

Progress in the field received further impetus with the development of synthetic steroids exhibiting activities far greater than that of the natural hormones. The synthesis of 9 α -fluorocortisol (3), described in 1953 by Fried and Sabo (3), opened the way for development of many more highly active antiinflammatory agents, and indeed some of today's most active antiinflammatory agents (ca 1997) bear some resemblance to this 9 α -fluorinated steroid.

Comprehensive reviews treating the pharmacological structure–activity relationships of glucocorticosteroids (4–7) and mineralocorticoids (8) have been published, as well as reviews of the mechanism of action (9,10). All natural adrenocorticoids are derivatives of the planar ring system 5 α -pregnane (4). Substituents lying above the plane of the rings are assigned a β -configuration, indicated by a dotted line. Angular methyl groups at C-10 and C-13 have the β -configuration and are often shown simply by solid bonds. Tertiary hydrogen atoms at C-8, C-9, C-14, and C-17 are usually omitted unless their stereochemistry differs from that shown in (4).



(4)

Clinical Use of Adrenal-Cortical Steroids

The predominant clinical use of corticosteroids is a result of their associated antiinflammatory properties. These are commonly used as topicals for the suppression of symptoms, including inflammation, occurring in a particular disease state; these compounds are rarely considered curative in their usage. Many other disease states do, however, respond well symptomatically to treatment with corticosteroid therapy. Some of these (11) are listed below.

	Antiinflammatory Activity
Adrenocortical insufficiency	Allergic conditions and conjunctivitis
Organ transplants	Cerebral edema
Liver disease	
Adrenogenital syndrome	GI diseases; ulcerative colitis and anorectal disorders
Nephrotic syndrome	Bacterial meningitis
Acute spinal cord injury	Rheumatic disorders and collagen diseases
Hypercalcemia	Ophthalmic, otic, and nasal disorders
Hematologic disorders	Dermatologic diseases
Myasthenia gravis	Respiratory diseases
Neoplastic disease	

Mineralocorticoid therapy is a less common, though still important aspect of the medicinal use of adrenal-cortical steroids. Very important applications include use in treating adrenocortical insufficiency or in other adrenocoritcoid replacement therapies, and in the management of salt-losing forms of congenital adrenogenital syndrome. Aldosterone, owing to its short half-life, is not used therapeutically; desoxycorticosterone acetate (a natural aldosterone precursor) and fludrocortisone are the only clinically significant mineralocorticoidal agents.

Topical Corticosteroids Currently Available. Table 1 includes compounds listed in *Drug Facts and Comparisons* (FC) (12) and the *AHFS Drug Index* (DI) (13). The compounds listed by FC as topical antiinflammatories were placed within a range of Low to Very High Potency (see Table 1). The DI list indicates a range from I (most active) to VI (least active), shown next to drug names listed in the table. Some compounds have more than one activity class, depending on the method of administration and concentration; in these cases, the higher class is listed. As has been noted elsewhere in the literature, activity ...

... may vary considerably depending on the vehicle, site of application, disease, the individual patient, and whether or not an occlusive dressing is used. The approximate relative activity is based principally on vasoconstrictor assay and/or clinical effectiveness in psoriasis (11).

Table 1. Topical Corticosteroid Compounds—*Drug Facts and Comparisons*^a vs *AHFS Drug Information*^b

Substance	Chemical name	Structure number	Level of activity ^c	Degree of potency ^d			
				Very high	High	Medium	Low
β-methasone dipropionate	9-fluoro-11β-hydroxy-16β-methyl-17α,21-bis(1-oxopropanoxy)-pregna-1,4-diene-3,20-dione		I	X			
clobetasol propionate	21-chloro-9-fluoro-11β-hydroxy-16β-methyl-17-(1-oxopropoxy)-pregna-1,4-diene-3,20-dione	(102)	I	X			
diflorasone diacetate	17,21-bis(acetyloxy)-6α,9-difluoro-11β-hydroxy-16β-methylpregna-1,4-diene-3,20-dione	(100)	I	X			
halobetasol propionate	21-chloro-6α,9-difluoro-11β-hydroxy-16β-methyl-17-(1-oxopropoxy)-pregna-1,4-diene-3,20-dione			X			
amcinonide	21-(acetyloxy)-16,17α-[cyclopentylidenebis(oxy)]-9-fluoro-11β-hydroxypregna-1,4-diene-3,20-dione		II		X		
β-methasone valerate	9-fluoro-11β,17α,21-trihydroxy-16β-methyl-17-[(1-oxopentyl)oxyl]-pregna-1,4-diene-3,20-dione	(99)	III		X		
desoximetasone	9-fluoro-11β,21-dihydroxy-16α-methylpregna-1,4-diene-3,20-dione		II		X		
fluocinonide	21-(acetyloxy)-6α,9-difluoro-11β-hydroxy-16,17-[(1-methylethylidene)bis(oxy)]-pregna-1,4-diene-3,20-dione		II		X		
fluocinolone acetonide	6α,9-difluoro-11β,21-dihydroxy-16α,17-[(1-methylethylidene)bis(oxy)]-pregna-1,4-diene-3,20-dione		IV		X		
halcinonide	21-chloro-9-fluoro-11β-hydroxy-16α,17-[(1-methylethylidene)bis(oxy)]-pregn-4-ene-3,20-dione		II		X		
mometasone acetonide			III				
triamcinolone acetonide	9-fluoro-11β,21-dihydroxy-16α,17-[(1-methylethylidene)-bis(oxy)]-pregna-1,4-diene-3,20-dione	(49)	III		X		
β-methasone benzoate	17-(benzoyloxy)-9-fluoro-11β,21-dihydroxy-16β-methylpregna-1,4-diene-3,20-dione		III			X	

HORMONES, ADRENAL-CORTICAL			Supplement
clocortolone pivalate	9-chloro-21-(2,2-dimethyl-1-oxopropoxy)-6 α -fluoro-11 β -hydroxy-16 α -methylpregna-1,4-diene-3,20-dione		X
flurandrenolide	6 α -fluoro-11 β ,21-dihydroxy-16 α ,17-[(1-methylethylidene)-bis(oxy)]-pregn-4-ene-3,20-dione	IV	X
fluticasone propionate	6 α ,9-difluoro-11 β -hydroxy-16 α -methyl-3-oxo-17 α -(1-oxopropoxy)-androsta-1,4-diene-17 β -carbothioic acid, 5-fluoromethyl ester		X
hydrocortisone butyrate	11 β ,21-dihydroxy-17-(1-oxobutoxy)-pregn-4-ene-3,20-dione	V	X
hydrocortisone valerate	11 β ,21-dihydroxy-17-[(1-oxopentyl)oxy]-pregn-4-ene-3,20-dione	V	X
prednicarbate mometasone fluroate	9 α ,21-dichloro-17-[(2-furanyl-carbonyl)oxy]-11 β -hydroxy-16 α -methyl-pregna-1,4-diene-3,20-dione	V	X
adometasone dipropionate	7 α -chloro-11 β -hydroxy-16 α -methyl-17 α ,21-bis(1-oxopropanoxy)-pregna-1,4-diene-3,20-dione	VI	X
desonide	11 β ,21-dihydroxy-16 α ,17-[(1-methylethylidene)bis(oxy)]-pregna-1,4-diene-3,20-dione	VI	X
dexamethasone	9-fluoro-11 β ,17,21-trihydroxy-16 α -methylpregna-1,4-diene-3,20-dione	(74)	X
dexamethasone sodium phosphate	9-fluoro-11 β ,21-dihydroxy-16 α -methyl-21-(phosphonoxy)-pregna-1,4-diene-3,20-dione, disodium salt		X
hydrocortisone	11 β ,17,21-trihydroxy-pregn-4-ene-3,20-dione		X
hydrocortisone acetate	21-(acetyloxy)-11 β ,17-dihydroxy-pregn-4-ene-3,20-dione		X

^a Ref. 12.

^b Ref. 13.

^c Ranging from I (most active) to VI (least active), as specified by the DI (13).

^d Range in which compounds listed by FC were placed (12).

Adrenocortical Biosynthesis and Metabolism

Corticosteroids are biosynthesized from cholesterol and released as needed by the adrenal cortex; they are not stored. Cholesterol undergoes a series of irreversible oxidations during which carbons 22 through 27 are cleaved, resulting in pregnenolone. Reversible isomerization of Δ^5 to Δ^4 results in progesterone, the key intermediate in both mineralocorticoid and glucocorticoid biosynthesis. Cortisol is the product of the 11 β -, 17 α -, and 21-oxidations (flavoprotein, cytochrom P-450-mediated) of progesterone; aldosterone is the result of oxidations at the 11 β -, 18-, and 21-positions. Glucocorticosteroid biosynthesis is outlined in Figure 1, and key steps in the mineralocorticoid biosynthesis are highlighted in Figure 2.

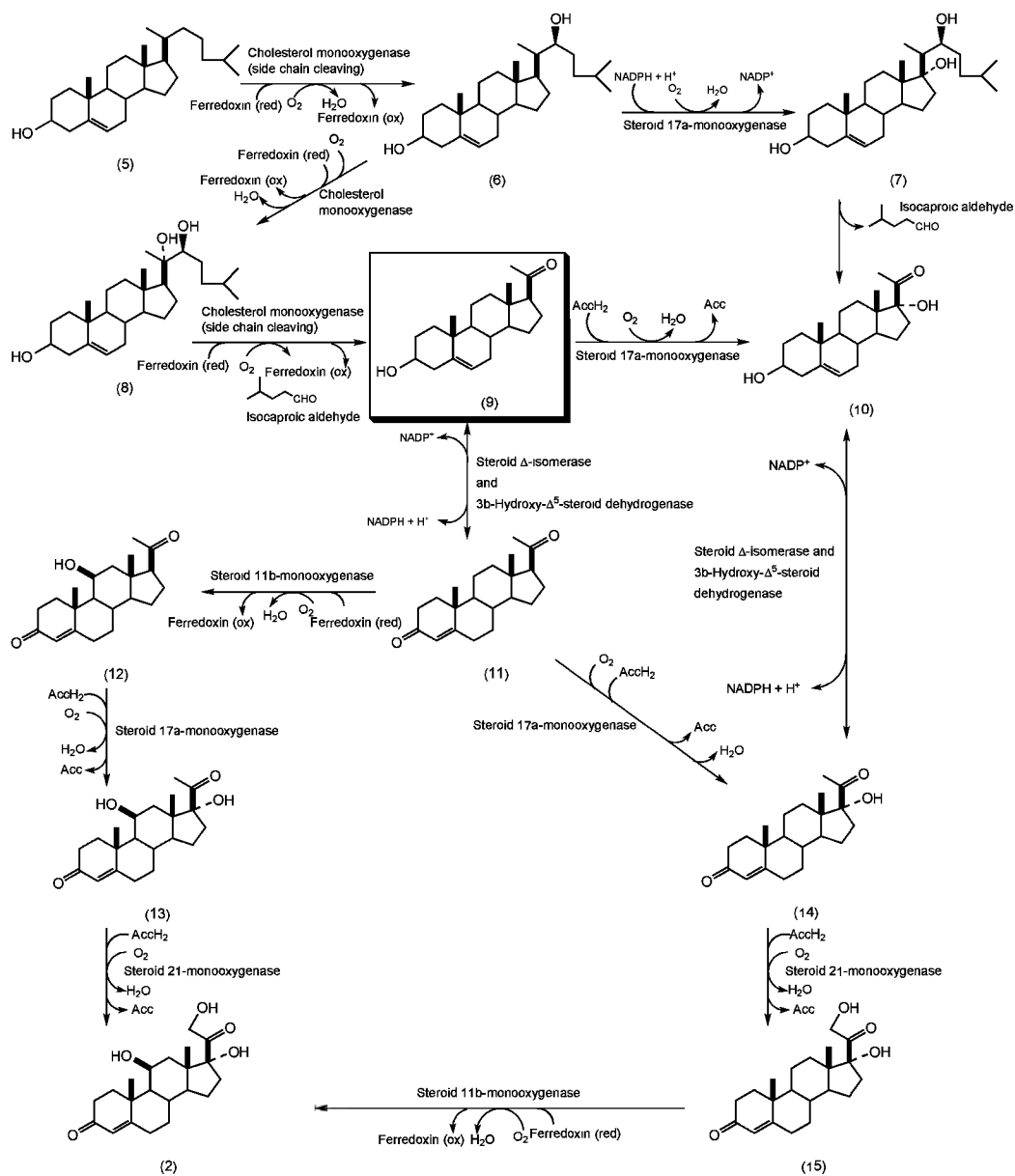


Fig. 1. Biosynthetic pathways for formation of cortisol from cholesterol.

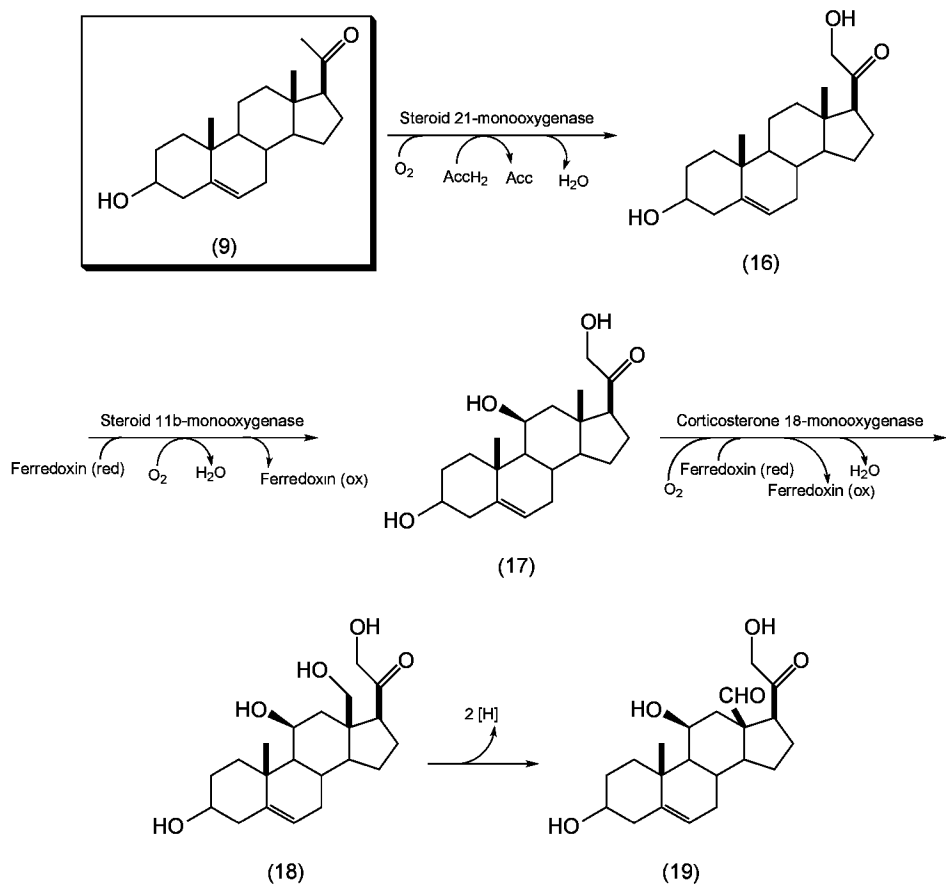


Fig. 2. Key steps in the mineralocorticoid biosynthesis.

The major metabolic transformations of the adrenal-cortical hormones generally follow the metabolism of cortisol, which undergoes the following conversions *in vitro*:

Cortisol-Cortisone Conversion. Under normal conditions, this equilibrium slightly favors the oxidized compound. Similarly, the conversion of corticosterone to 11-deoxycorticosterone is also mediated by the 11 β -hydroxysteroid dehydrogenase enzyme system and requires NAD(P)⁺ NAD(P)H. This conversion is especially important both in the protection of the human fetus from excessive glucocorticoid exposure, and in the protection of distal nephron mineralocorticoid receptors from glucocorticoid exposure (14). The impairment of this conversion is thought to result in hypertension associated with renal insufficiency (15).

A-Ring Reduction. This is an irreversible reaction which is a foremost determinant of the secretion rate of cortisol (double bonds and C-3 carbonyl). Catalyzed predominantly by cortisone β -reductase and 3 α -hydroxysteroid dehydrogenases, 5 β sterols result, although 5 α sterols are more prevalent in the case of other glucocorticoids. Urocortisol and urocortisone result from the metabolism of cortisol and cortisone, respectively. Compounds can be complexed to glucuronic acid at this point.

C-20 Reduction. Two stereoisomers can result from this transformation, although cortisol is thought to act primarily with (R)20 β -hydroxysteroid dehydrogenase. This is a first step in the metabolism of corticosterone.

Cleavage of C-17 Acyl/Alkyl Substituents. Resulting primarily in cholan-17-ones, this is a relatively minor metabolic pathway. Corticosterone is not known to undergo this transformation before excretion.

C-6 Hydroxylation. This biotransformation is more predominant in infants than in adults, and can prevent other metabolic transformations.

Glucuronidation. Complexation of the steroid to glucuronic acid, most predominantly via the C-3 hydroxyl, leads to a considerable portion of the excreted metabolites of all glucocorticoids. In infants, sulfurylation (formation of a sulfate ester) is also predominant (16).

Other Reactions. Most of the metabolites of cortisol are neutral (alcohol or glucuronide complex) compounds. However, oxidation at C-21 to C-21 carboxylic acids (17) accounts for some of the identifiable metabolites of glucocorticoids (18).

Compounds having the 16,17 ketal, eg, budesonide, amcinonide, fluocinonide, halcinonide, triamcinolone acetonide, and flurandrenolide, also undergo metabolism by routes that parallel that of cortisol metabolism. Unsymmetrical acetals such as budesonide are also metabolized by routes not available to the more metabolically stable symmetrical 16 α ,17 α -isopropylidene-dioxysubstituted compounds (desonide, flunisolide, and triamcinolone acetonide). Isozymes within the cytochrome P450 3A subfamily are thought to catalyze the metabolism of budesonide, resulting in formation of 16 α -hydroxyprednisolone and 6 β -hydroxybudesonide (19,20) (Fig. 3) in addition to the more common metabolic steps (oxidation via Δ^3 , reduction of Δ^3 , etc).

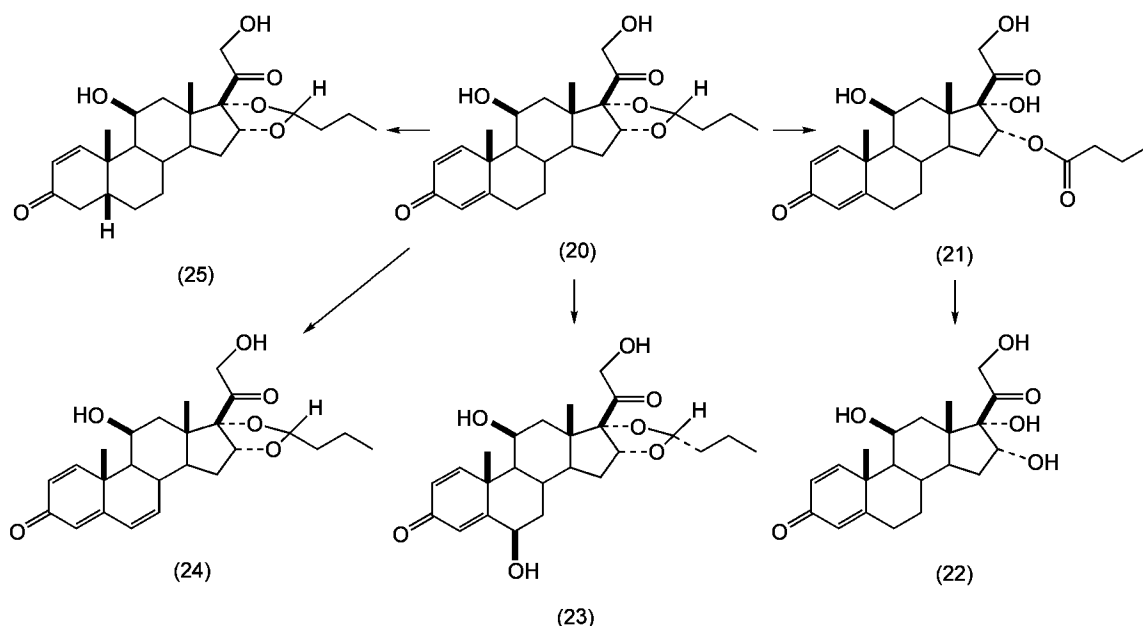


Fig. 3. Metabolism of 22R-budesonide in human liver microsomes.

Steroids having the greatest number of substituents generally have the slowest rate of metabolism. Groups that appear to activate by effects on metabolism include 2 β -methyl, which stabilizes the resulting molecule to the action of the 4,5-reductase (21) and to the action of 20-keto reductase (22). Similarly, 6 α -methyl protects the A ring against metabolic destruction (22). Again, the introduction of 16 α -hydroxyl, as in triamcinolone (48), prolongs the half-life (23). The 16 α -methyl and 16 β -methyl groups have similar action. Triamcinolone is unusual in that it is metabolized principally to the 6 β -hydroxy derivative (23).

Mechanism of Action of Antiinflammatory Steroids

Receptor Structures for Glucocorticoids and Mineralocorticoids. The effects of adrenal-cortical steroids are thought to result from their interaction with intracellular receptors, and a great deal of attention has been given to the task of determining the structure and function of the glucocorticoid receptor (GR) as well as the mineralocorticoid receptor (MR). Though studied in less detail, the primary function of the MR seems to be nearly identical to that of the GR, with the primary differences being the restricted expression of the MR (limited mostly to tissues in the kidney, colon, salivary and sweat glands, and hippocampus), and the different proteins encoded by the activated DNA-receptor complex. The highly homologous GR and MR proteins are also very similar in action, and the description of the more thoroughly examined GR is adequate for the understanding of the MR structure and function.

The primary structure of the GR has been identified as a 795 residue protein. Though there does not yet exist an x-ray structure, a significant amount of 3-D structure elucidation of the GR has been made. The GR is a member of the nuclear receptor superfamily, which includes receptors for other steroids, vitamin D, and thyroid hormones, and some other various proteins (24). A high degree of homology is found between receptors in this class, and each contains a similar domain make-up. These domains (sections of primary structure) each have a functional duty, and generally describe regions of hormone binding, nuclear translocation, dimerization, DNA binding, and transactivation (25).

Within the carboxyl terminus portion of the protein (residues 518–795) is the ligand (hormone) binding domain (HBD) (25). It is here that much of the interaction of the receptor with hsp90 occurs (26) (thought to be residues 595 to 614 of the rat GR (27)). Three of five cystein residues in this area are spaced close together in the binding pocket (28), and two of these are critical to the receptor's ability to bind specifically glucocorticoids (29). Binding of the hormone to the HBD allows the activation of the receptor to the DNA-binding state, and during this process, the conformation of the protein changes, in part the result of the dissociation of hsp90. Antigluocorticoids and certain other modulators as well as sodium molybdate will interact with this domain

and inhibit activation.

The DNA binding domain is highly conserved among species, and changes to the amino acid sequence in this region result in changes in receptor function (30). A structural feature which characterizes the GR-DNA binding domain are the two so-called zinc fingers, in each of which two zinc (+2) ions are held in place by tetrahedral coordination to neighboring cysteine residues (31). These zinc fingers, common to the nuclear receptor superfamily, are not exactly the same as those found in *Xenopus TFIIIA*, in which histidine residues are also associated with the metal (32,33). The zinc in the GR is thought to stabilize the α -helices at the carboxy ends of the fingers as well as aid in carboxy-terminal module folding, needed in the dimerization of the proteins (34). Though the entire glucocorticoid receptor (GR) tertiary structure has yet to be elucidated, some of the components of the structure, including the DNA-binding domain, have been depicted through molecular modeling studies (35), nmr investigations (36), and the crystal structure data from GR protein fragments (34). Solution-state analysis of the DNA-binding domain complexed to a nucleotide sequence (double helix) reveals areas where an α -helical substructure interacts with the nucleotide. This DNA-receptor complex structure is supported by the crystal structure of a similar DNA-receptor fragment (DBD) complex. Clear dimeric interaction can be seen, along with the general shape of the zinc finger region.

Two domains, τ 1 and τ 2, exist which affect the GR post-DNA binding transcription activity (37). The major (τ 1) transactivation domain is 185 amino acid residues in length with a 58-residue α -helical functional core (38). The τ 1 domain is located at the N terminus of the protein; the minor (τ 2) transactivation domain residues on the carboxy-terminal side of the DNA binding domain.

Mechanism of Action. The process by which adrenocortical steroids impart their action is based on the action of the steroid on a receptor (MR or GR). One result of this process is the lag between optimum pharmacologic activity and peak blood concentrations. The stages of GR goes through in becoming active can be divided into five general steps (39), and each step is mediated by the glucocorticoid receptor (GR).

(1) *Subcellular Localization.* Some debate still exists (ca 1997) as to whether GRs are cytoplasmic or nuclear in nature (40,41), though it is believed that the interaction of hormone and receptor occurs in the cytoplasm. Compounds with high binding affinities (glucocorticoid agonists or antagonists) result in complete translocation of these receptors into the nucleus (42). Phosphorylation sites on the GR do not seem to play a role in hormone-inducible nuclear translocation. Sequences controlling nuclear localization have been identified within steroid receptors in the hinge region. The glucocorticoid receptor has a second nuclear localization signal in the hormone binding domain (43).

(2) *Association with Heat Shock Proteins (hsp).* Unactivated GRs are complexed with a number of protein factors which play various roles in the binding of ligand to the receptor, as well as the localization, DNA binding, and transactivation of the GR. The proteins, including hsp90, hsp70, hsp56, CyP40 and p23 (an acidic protein), have been implicated as part of an assembled complex that is able to activate GR to the ligand binding state; this complex is termed a foldosome (44,45). Hsp90 has been shown to be associated with the ligand-binding portion (C terminus) of the receptor (26), perhaps even blocking this site, though it is needed for the GR ligand-binding domain to be in the proper conformation for ligand binding (46). Hsp70 assists the binding of hsp90 to the receptor. Upon ligand binding, the heat shock proteins dissociate and the receptors become active in dimerization, DNA binding, and transcriptional enhancement (47). Though these hsp proteins do block certain DNA binding sites on the receptor, protein-free receptor studies have indicated that the free receptor still needs to be bound to hormone in order to bind to DNA.

(3) *Hormone binding.* The ligand-binding domain of the steroid hormone receptors encompasses almost the entire C terminus of these proteins. This domain is also credited with having the functions of hormone-dependent dimerization and transactivation. The ligand-binding domain of the glucocorticoid receptor appears to repress transcription in the absence of hormone, and this transrepression is reversed by hormone. Active glucocorticoid receptor agonists bind tightly to the hormone receptor to elicit their action. Binding of glucocorticoids to the GR hormone-binding domain transforms the receptor into an activated complex able to interact with DNA sequences called the glucocorticoid response elements (GREs) of target genes. One difference between the MR and GR is in the affinities each has for cortisol and aldosterone. Both of these steroids bind to the MR (type I corticosteroid receptor), and cortisone also binds to the GR (type II corticosteroid receptor). The enzyme 11 β -hydroxysteroid dehydrogenase plays an important role in converting cortisol to a less active cortisone, thus allowing aldosterone to bind freely to the MR.

(4) *Dimerization and DNA binding.* The active regulatory form of the GR is thought to be dimeric. The GR binds to a palindromic DNA sequence (GRE), either as a GR dimer (48), or perhaps as a GR monomer followed by binding of a second GR. Crystallographic studies with GR fragments containing the DNA-binding domain have indicated that this dimerization occurs upon binding to the half-site of the GRE (34). Members of the steroid hormone receptor superfamily contain a highly conserved DNA-binding domain of about 70 residues, which are complexed around two tetrahedrally coordinated zinc atoms.

(5) *Transactivation.* Protein synthesis is initiated or inhibited by the action of the activated GR on DNA. The use of glucocorticoids leads to antiinflammatory effects by first controlling gene expression, which subsequently leads to the synthesis and or suppression of inflammation regulatory proteins.

One such regulatory protein, inducible by a number of glucocorticoids, is the 37kDa protein Lipocortin 1 (LC-1) (49,50). Dexamethasone directly effects *de novo* LC-1 synthesis, leading to direct increases in intra- and extracellular LC-1 concentrations, occurring most notably at the cell surface (51). Prednisolone increases extracellular and decreases intracellular concentrations (52) of LC-1. Studies with LC-1 and with an active LC-1 segment show direct involvement of this protein in inhibiting neutrophil activation via inhibition of the release of elastase, PAF, leukotriene B4, and arachidonic acid (53), as well as an inhibition of neutrophil adhesion to endothelial monolayers (54,55). LC-1 inhibits various types of prostanoid (inflammation mediator) production; it also suppresses thromboxane A₂ release from perfused lungs (56) and has thus been shown to inhibit inflammation in the rat paw (carragenen-induced edema) inflammation model (57). This is a result of LC-1 inhibition of phospholipase A₂, which converts membrane phospholipids to arachidonic acid, along with its effect on other cellular components of certain inflammatory responses (10,58). Antibodies raised against LC-1 are able to reverse the antiinflammatory action of LC-1 (59–61). Corticosteroids reduce phospholipase A₂ activity, which results in the diminished release of arachidonic acid, and this subsequently leads to limiting the formation of prostaglandins, thromboxane, and the leukotrienes (9,62).

Glucocorticoids have been shown to inhibit gene transcription of other proteins involved in the inflammatory process, including the key inflammation mediators called cytokines (IL-1, IL3–6, IL8, GM-CSF, TNF α) (10,58,63–65). Steroids have been also shown to suppress the formation of cytokine receptors (10); dexamethasone, in particular, downregulates gene transcription of angiotensin II type 2 receptors (66).

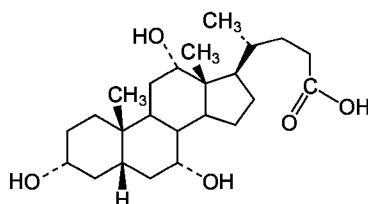
Mineralocorticoids follow a mechanistic route similar to that of glucocorticoids, though differing in the proteins expressed. The activated MR-DNA complex promotes the expression of aldosterone-induced proteins (AIPs), which then act to increase Na⁺ conductance of the luminal membrane and concurrently increase Na⁺ pump activity of the basolateral membrane. These actions result from a number of AIP-influenced cellular characteristics, including a modification of tight junction permeability, an increase in membranous Na⁺ channels and pumps, activation of silent Na⁺ channels and pumps, and an increase in mitochondrial ATP-production-related enzymes.

Synthesis of Glucocorticoids

Hydrocortisone and Prednisolone. Following the discovery of the antiinflammatory actions of cortisone (1) and cortisol (2), there was a need not only to develop highly efficient routes to the corticoids, but to discover novel structures with fewer side effects than those of the corticoids, eg, sodium and water retention, reduced carbohydrate tolerance (steroid diabetes), osteoporosis, and depressed host defense.

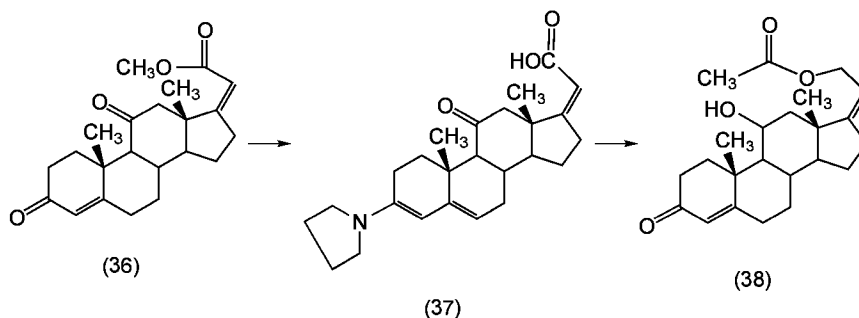
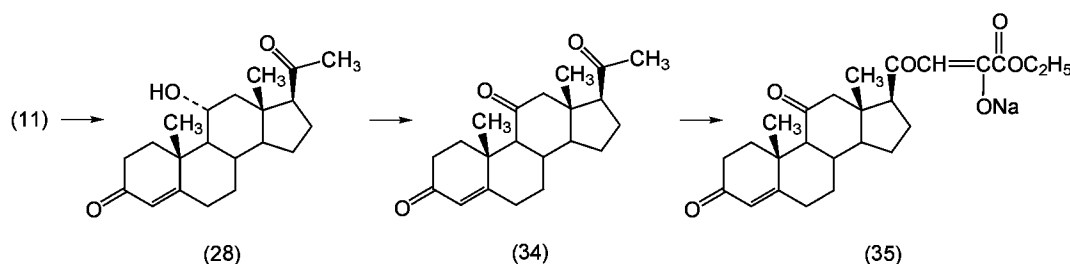
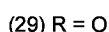
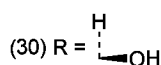
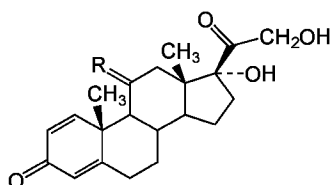
A major difficulty in the manufacture of corticosteroids was the lack of an abundant raw material containing an 11-oxygenated function. This problem initially was solved by using 12 α -hydroxylated cholic acid (26) (obtained from ox bile); later it was converted into cortisone acetate by a 32-stage process (66). A more recent development in procuring readily available starting material was the discovery that perfusion of cortexone (16) through the adrenal gland gives corticosterone (17), albeit in low yield, thereby demonstrating the existence of enzymes capable of hydroxylating progesterone derivatives at position C-11. Microorganisms capable of introducing an 11 β -hydroxyl group into a steroid were discovered. *Rhizopus arrhizus* transformed progesterone (11)—which was readily available from diosgenin or from the soybean sterol, stigmasterol (27)—into 11 α -hydroxyprogesterone (28) in excellent yield; even better results were obtained using *Rhizopus nigricans* (67). Other organisms include *Corynebacterium simplex* which converts cortisone (1) and cortisol (2) into their 1-dehydro derivatives, prednisone (29) and prednisolone (30) (68), respectively. These steroids surpassed their parent hormones

in antirheumatic and antiallergic activity and produced lower mineralocorticoid activity and other side effects. Nearly all corticoids on the market other than cortisone are 1-dehydro steroids.



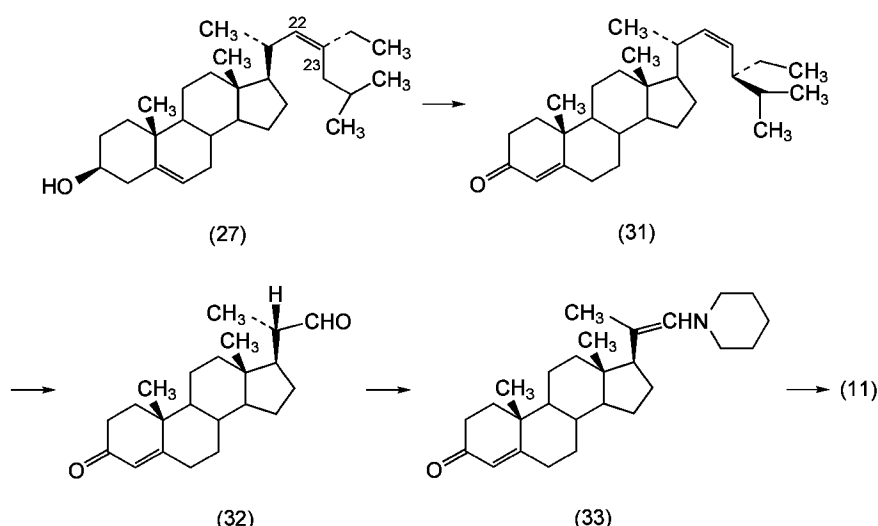
(26)

The classic process of corticosteroid manufacture employs stigmasterol (27), obtained from soybean oil, as the raw material. The sterol is first oxidized to stigmasteradien-3-one (31) by the Oppenauer method. Ozonolysis of the latter in methylene chloride 1° o pyridine yields 3-oxobisnorchol-4-enyl aldehyde (32) (69). Azeotropic distillation of the aldehyde with piperidine containing a trace of *p*-toluenesulfonic acid (pTS) leads to formation of the enamine (33) (70) which, when oxidized with sodium dichromate in anhydrous acetic acid, gives progesterone (11) in high yield. Microbiological oxidation of the (resulting) progesterone (11) with *Rhizopus nigricans* yields 11 α -hydroxyprogesterone (28) in yields exceeding 90% (67). The latter is oxidized to 11-oxoprogesterone (34), which is condensed with diethyl oxalate in the presence of sodium methoxide to yield the 21-ethoxyoxalyl derivative (35). The last compound is treated with bromine giving the 21,21-dibromo-21-carbethoxylate, which is treated without isolation



with sodium methoxide whereby it undergoes Favorski rearrangement to give the methyl-*cis*-pregn-17-enoate derivative (36). Protection of the 3-oxo- Δ^4 -system in (36) (as the enamine (37) permits reduction of the 11-oxo- and carbomethoxy groups with diisobutyl aluminum hydride or LiAlH₄ in dibutyl ether. Regeneration of the 3-oxo function and acetylation gives (38). Oxidation of the latter with phenyliodosoacetate in *t*-butanol in the presence of a catalytic amount of OsO₄ gives hydrocortisone acetate (2, acetate) (71).

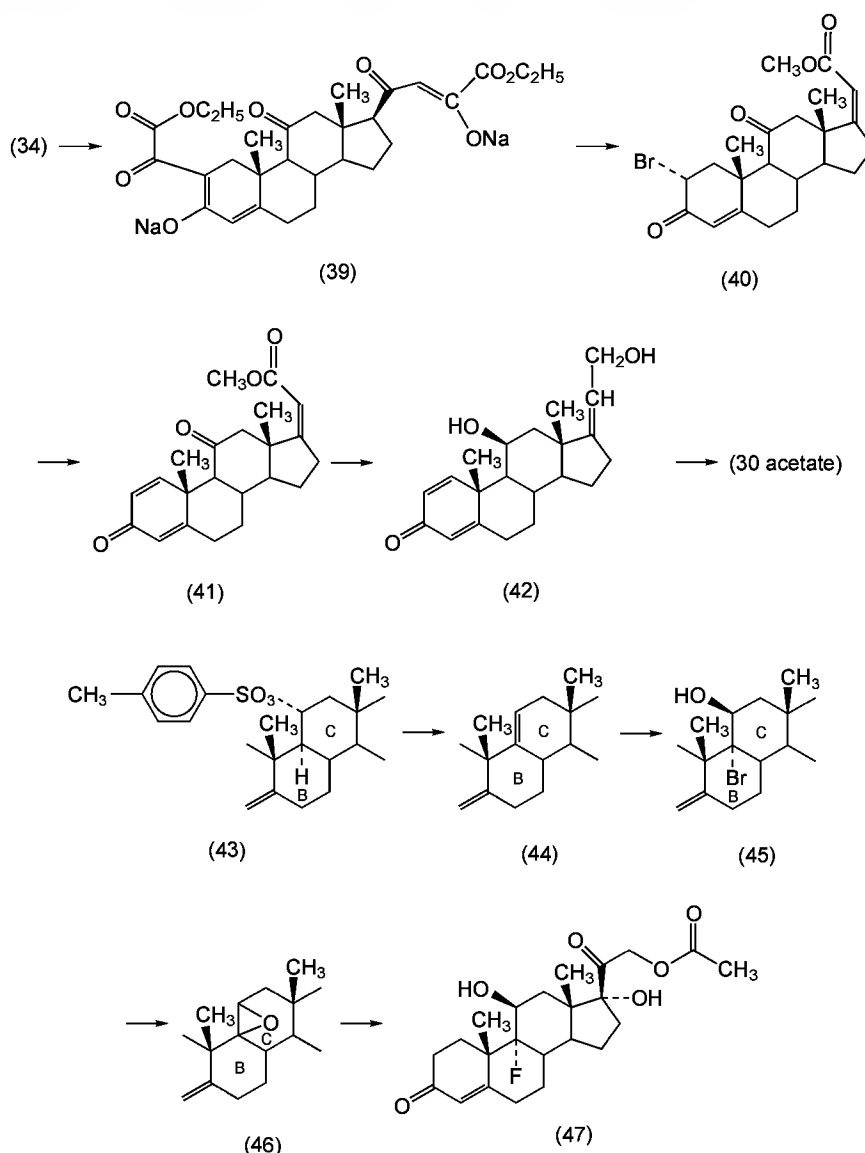
The manufacture of prednisolone (30) follows a similar route. 11-Oxo-pregesterone (34) is treated with two molar equivalents of diethyl oxalate and sodium methoxide in *t*-butanol. The product (39) is acidified and treated with



three moles of bromine NaOCOCH_3 , $\text{CH}_3\text{CO}_2\text{H}$ to yield a tribromo derivative which undergoes Favorski rearrangement on treatment with NaOCH_3 , CH_3OH to give the 2-bromopregnenic acid derivative (40). Dehydrobromination with Li_2CO_3 in refluxing dimethylformamide yields the 1,4-diene (41), which is treated with dimethylamine in the presence of titanium chloride (72) to yield the 3-enamine. The 3-enamine is treated to form the 3,11-dienamine which can be converted into prednisone (29). Reduction with diisobutylaluminum hydride and regeneration of the 3-oxo moiety furnishes (42), which is converted after acetylation by hydrogen peroxide (with OsO_4 as catalyst) into prednisolone 21-acetate (30-acetate).

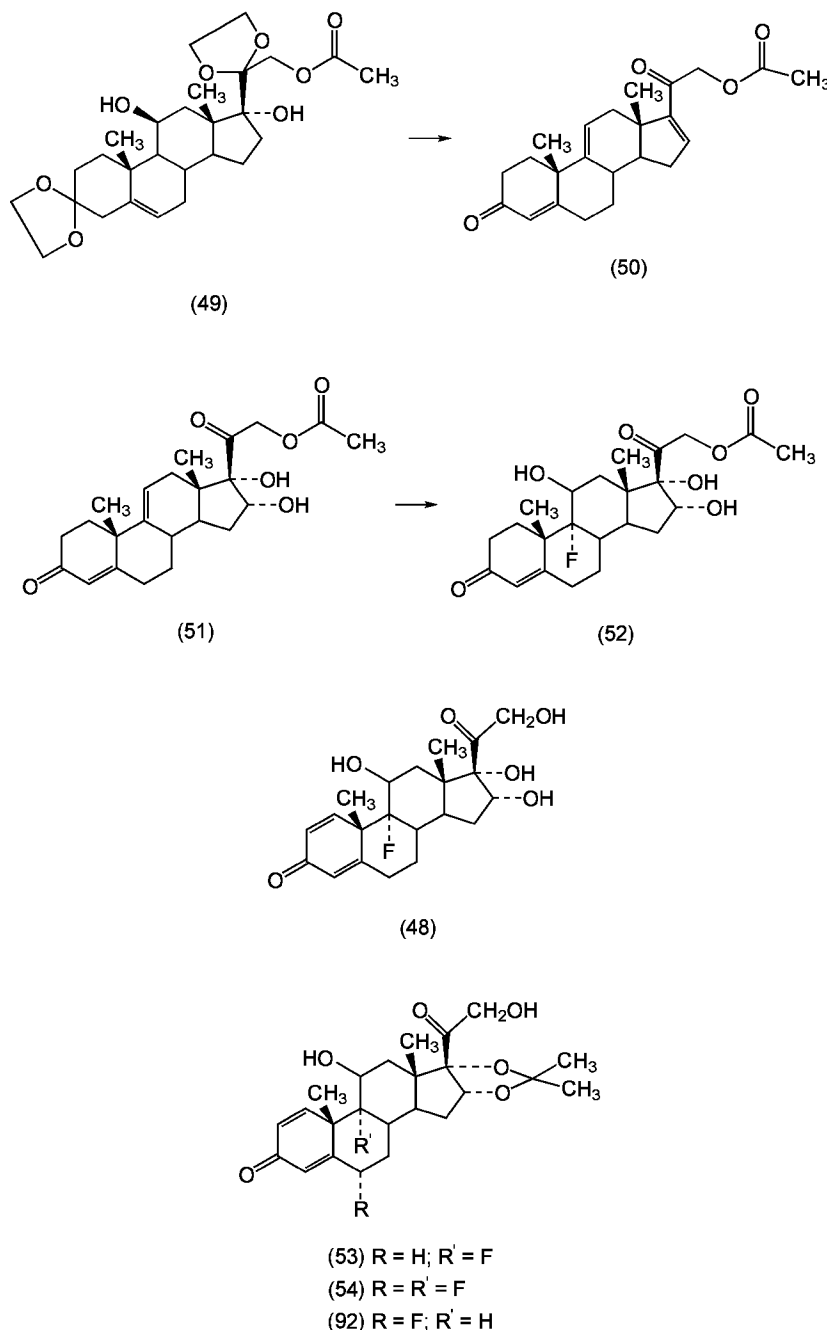
9-Fluoro Derivatives of Corticoids. The 11-tosylate (43) of 11-epicortisol 21-acetate has been treated with sodium acetate-acetic acid in the hope of enforcing epimerization of the 11α -hydroxyl group. Surprisingly, the 9,11-ene (44) was formed by cis-elimination of the tosyl group. Addition of hypobromous acid gave the bromohydrin (45), which was converted by sodium acetate into the $9\beta,11\beta$ -epoxide (46). The latter was treated with hydrofluoric acid to give 9α -fluorohydrocortisone acetate (fludrocortisone acetate) (47). When tested in patients with rheumatoid arthritis, fludrocortisone acetate (47) proved to be 10 times more potent than hydrocortisone acetate. In contrast to hydrocortisone, which has only moderate mineralocorticoid activity, fluorocortisone acetate produces edema through salt and water retention. However, it has found clinical use in the treatment of patients with Addison's disease. Virtually every corticoid analogue prepared subsequent to (47) was routinely converted into the 9α -fluoro derivative, and nearly all of the corticoids in use contain this moiety.

16 α -Hydroxy Derivatives of Corticoids and their Acetonides. The preparation of 16 α -hydroxy-9 α -fluoroprednisolone (48) from the 3,20-bisethylene ketal of hydrocortisone acetate (49) has been reported (73). The latter was dehydrated with thionyl chloride in pyridine to yield the 4,9(11),16-triene (50). The 16,17-unsaturated linkage was selectively hydroxylated with OsO_4 pyridine to yield the 16 $\alpha,17\alpha$ -diol (51), which was converted into the 9α -fluoro-11 β -hydroxy steroid (52) by the procedure outlined in (44) — (47). Microbiological oxidation with *Corynebacterium simplex* gave the 1-dehydro derivative, triamcinolone (48). Triamcinolone possesses high glucocorticoid and antiinflammatory

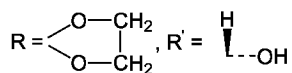


activity and, because of the presence of the 16 α -hydroxyl group, is almost entirely devoid of salt-retaining activity. It had been shown that steroidal 16 α ,17 α -dihydroxy-pregnan-20-ones readily forms acetonides, and application of this reaction to triamcinolone (48) yields triamcinolone acetonide (53) (74) which significantly exceeds the parent steroid in topical antiinflammatory activity. 16 α -Hydroxyprednisolone 16 α ,17 α -acetonide (desonide) (54) has been introduced into clinical practice (75).

Methylated Glucocorticoids. The preparation of 2 α -methyl-9 α -fluorocortisol has been reported (76). This compound shows enhanced glucocorticoid activity and greatly enhanced mineralocorticoid activity, so much that it surpasses aldosterone (19) in sodium-retaining and potassium-excreting potency. Attention was then turned to the preparation of 6-methylated corticoids



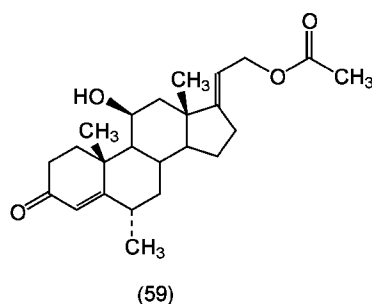
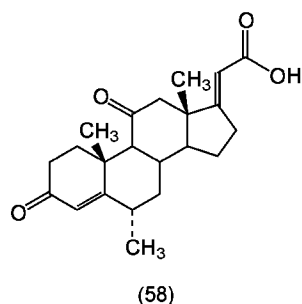
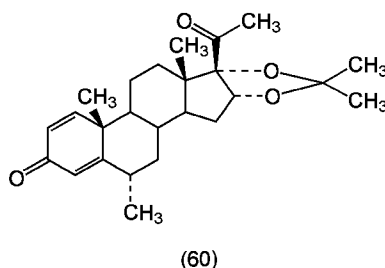
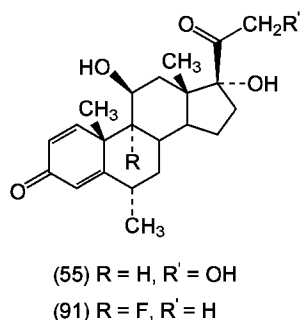
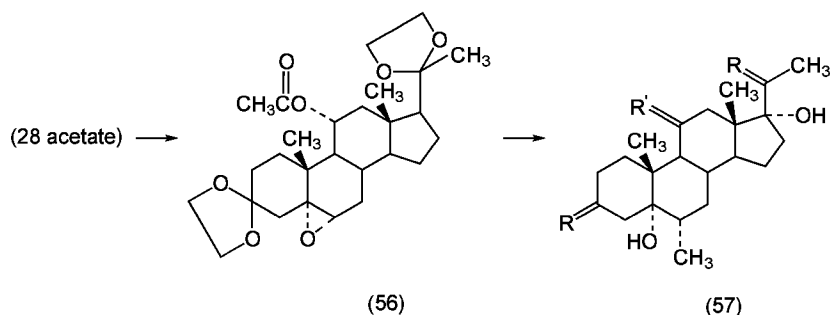
when potentiation of glucocorticoid activity linked to negligible mineralocorticoid activity resulted: 6 α -methylprednisolone (55) proved to be the compound of choice. It was prepared from 11 α -hydroxyprogesterone acetate (28-acetate), which was converted into the bis-ethylene ketal and then by reaction with peracetic acid into the 5 α ,6 α -epoxide (56). The latter compound was treated with methyl magnesium bromide (when fission of the epoxide ring occurred) to give the 5 α -hydroxy-6 β -methyl-bis-ketal (57)



Removal of the ketal protecting groups, followed by oxidation with sodium dichromate in acetic acid gave 5 α -hydroxy-6 β -methyl-5 α -pregnane-3,11,20-trione (57) (R = R' = O). The latter was submitted to oxalylolation, dibromination, Favorski rearrangement, and zinc dust reduction (see (39) $\rightarrow$ (40)) to yield 3,11-diketo-6 α -methylpregna-4,17(20)-dien-21-oic acid (58) (see also (36)). The last compound was converted into the pyrrolidyl enamine (cf (37)) which, on reduction with LiAlH₄ and regeneration of the 3-keto group by treatment with buffered methanol followed by acetylation, gave 11 β ,21-dihydroxyl-6 α -methylpregna-4,17(20)-dien-3-one acetate (59). Oxidation with phenyliodosoacetate in *t*-butanol pyridine in the presence of catalytic OsO₄ furnished 6 α -methylhydrocortisone, which subsequently was converted into 6 α -methylprednisolone by microbiological oxidation with *Septomyxa affinis* (77). The potentiating effect of the 6 α -methyl group upon glucocorticoid activity is well illustrated by the observation that 16 α ,17 α -isopropylidene-6-methylpregna-1,4-diene-3,20-dione (60) (77) is an excellent topical antiinflammatory agent even though it lacks the 11 β -hydroxyl group characteristic of the natural glucocorticoids (78) and originally believed to be essential for glucocorticoid activity. The 21-desoxycorticoid 11 β ,17 α -dihydroxy-9 α -fluoro-6 α -methylpregna-1,4-diene-3,20-dione (fluoromethalone) (61) also is an excellent topical antiinflammatory steroid (79).

Methylation of hydrocortisone prednisolone in positions C-4, C-7, C-12, and C-21 failed to give useful products. Methylation at C-16, in contrast, led to 16 α - and 16 β -methyl-9 α -fluoroprednisolones which were exceptionally useful. Both series were prepared using 3 α -acetoxy-5 α -pregn-16-ene-11,20-dione derived from desoxycholic acid (80). A much shorter route was subsequently developed from

16-dehydropregnenolone (62) which was readily available from diosgenin.



Pregnadienolone (62) was converted into 16 α -methylpregnenolone (63) by reaction with $\text{CH}_3\text{MgI} \cdot \text{Cu}_2\text{Cl}_2$ (81). Catalytic hydrogenation of the last compound gave the 5 α -pregnane derivative (64) which passed smoothly on enforced acetylation into the enol acetate (65). The latter, on treatment with peracetic acid, yielded the epoxide (66), which was converted into 17 α -hydroxy-16 α -methyl-5 α -pregnan-20-one (67). Bromination of the last compound at C-21 gave (68) from which, by metathesis with acetate ion, 21-acetoxy-3 β ,17 α -methyl-5 α -pregnan-20-one (69) was obtained. Chromic acid oxidation of the last compound yielded the 3-one (70), which was converted using the 2,4-dibromo derivatives (71) into the 1,4-dien-3-one (72). Microbiological hydroxylation furnished the 11-epi 16 α -methyl prednisolone (73) converted by standard procedures into 9 α -fluoro-16 α -methyl-11 β ,17 α -trihydroxypregna-1,4-diene-3,20-dione (9 α -fluoro-16 α -methylprednisolone; dexamethasone) (74) (82). The last compound is used in the diagnosis of Cushing's syndrome.

Introduction of a 16 β -methyl group into the corticosteroid molecule was effected by a reaction (83) whereby a 16-dehydropregnenolone (62) was treated with diazomethane to form the pyrazoline (75) which was decomposed with perchloric acid in acetone to give the 16-methylpregn-16-en-20-one derivative (76). Catalytic hydrogenation yielded the 16 β -methyl intermediate (77), which was converted into 9 α -fluoro-16 β -methyl-11 β ,17 α ,21-trihydroxypregna-1,4-diene-3,20-dione (9 α -fluoro-16 β -methylprednisolone, betamethasone) (78) by standard procedures (80).

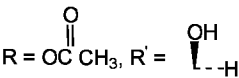
Hecogenin has been used as starting material for the preparation of betamethasone (78); this genin is found in the sisal plant, *Agave sisalana*. The 12-oxo group present in hecogenin (79) is transferred to C-11 (84). Bromination of hecogenin in benzene yields the 11 α ,23 α -dibromo derivative (80). The latter is treated with sodium hydroxide in aqueous *t*-butanol to yield the crystalline 23-bromo-ketal (81) which is acetylated and then debrominated with

zinc dust. Reduction with calcium liquid ammonia yields 11-oxotigogenin (82), which is converted into betamethasone (78) by standard procedures.

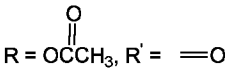
A variant of the 16-methyl group is 9α-fluoro-16-methylene-11β,17α,21-trihydroxypregna-1,4-diene-3,20-dione (fluprednylene) (83) (85).

6-Fluorocorticoids. The 6α-fluoro substituent resembles the 6α-methyl group in strongly potentiating glucocorticoid activity.

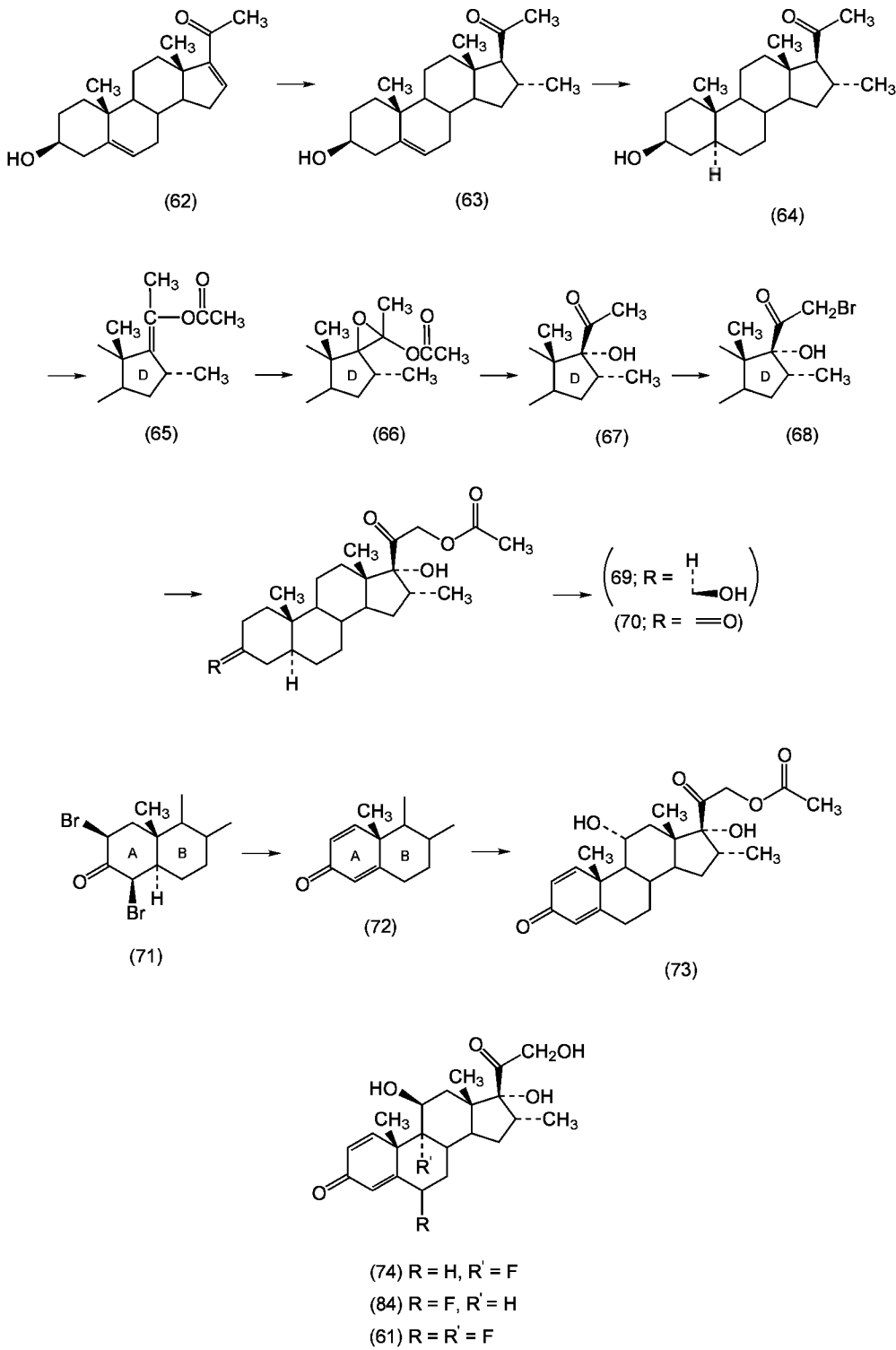
6α-Fluoro-16α-methyl-prednisolone (paramethasone) (84) was prepared from 16α-methylpregnenolone (acetate) (63). The 5α,6α-epoxide (85) of the last compound was treated with boron trifluoride to yield the 6β-fluoro-5α-hydrin (86) which gave the enol triacetate (87) with acetic anhydride acetyl chloride. Reaction with perphthalic acid gave the 17α-hydroxy-20-one (88) which was converted into the 21-acetoxy-derivative (89) by the standard bromination acetoxylation procedure. Oxidation

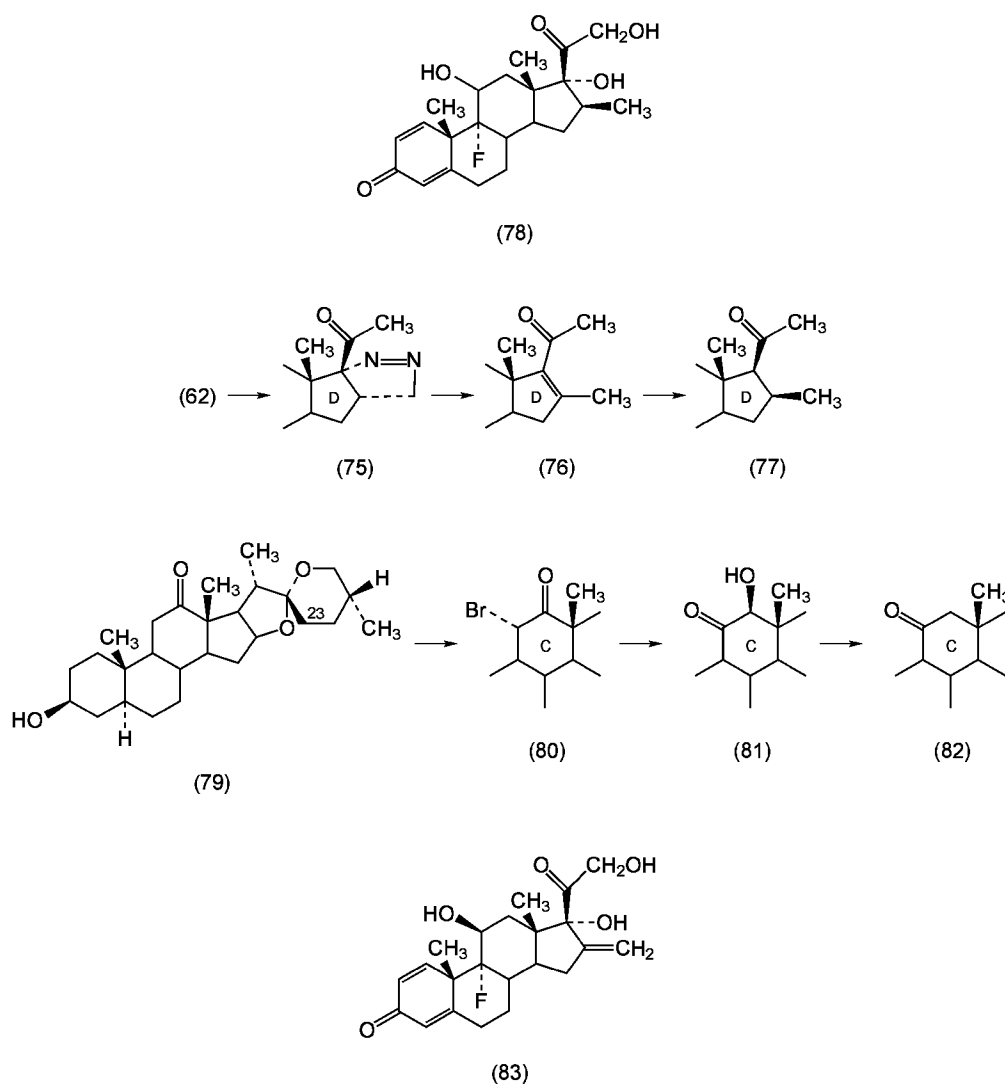


to the 3-one (89), followed by treatment with hydrogen chloride and cautious



alkaline hydrolysis gave 17α,21-dihydroxy-6α-fluoro-16α-methylpregn-4-ene-3,20-dione (90), converted into 6α-fluoro-16α-methyl-11β,17α,21-trihydroxy-





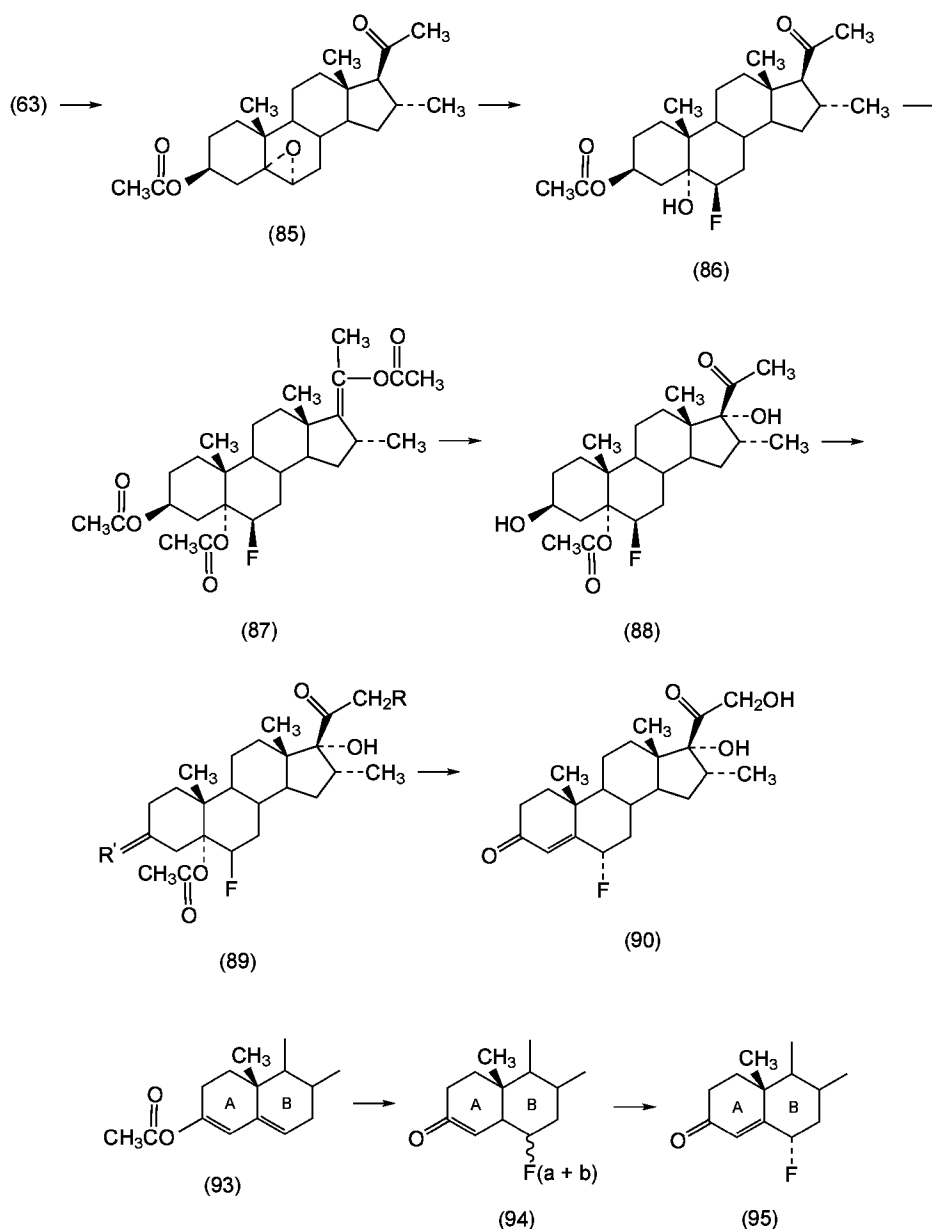
pregna-1,4-diene-3,20-dione (*para*-methasone) (84) and into 6 α ,9 α -difluoro-16 α -methyl-11 β ,17 α ,21-trihydroxypregna-1,4-diene-3,20-dione (flumethasone) (91) by standard procedures (86). The same general procedure was used to prepare

6 α ,9 α -difluoro-11 β ,16 α ,17 α ,21-tetrahydroxypregna-1,4-dien-3,20-dione-16,17-acetonide (fluocinolone acetonide) (54) (87), which proved to be extremely effective by the topical route and is widely used as a standard for determining the potency of new corticoids.

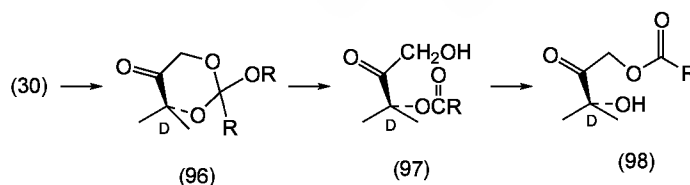
6 α -Fluoro-11 β ,16 α ,17 α ,21-tetrahydroxypregna-1:4-diene-3,20-dione-16 α ,17 α -acetonide (flurandrenolone acetonide) (92) also was prepared (88).

An alternative route to 6 α -fluorosteroids was developed by treating the enol acetate (93) of a Δ^4 -3-one with perchloryl fluoride in aqueous dioxane to obtain a mixture of the 6-fluoro-epimers (94) which were converted into the 6 α -fluoro- Δ^4 -3-ones (95) on treatment with H⁺.

17-Acylated Corticoids. The corticoid side-chain of (30) was converted into the cyclic ortho ester (96) by reaction with a lower alkyl ortho ester RC(OR')₂ in benzene solution in the presence of *para*-toluenesulfonic acid (88). Acid hydrolysis of the product at room temperature led to the formation of the 17-monoesters (97) in nearly quantitative yield. The 17-monoesters (97) underwent acyl migration to the 21-monoesters (98) on careful heating with H⁺. In this way, prednisolone 17 α ,21-methylorthovalerate was converted quantitatively into prednisolone 17-valerate, which is a very active antiinflammatory agent (89). The intermediate ortho esters also are active. Thus, 17 α ,21-(1'-methoxy)-pentylidenedioxy-1,4-pregnadiene-11 β -ol-3,20-dione [(96), R' = CH₃, R = C₄H₉] is at least 70 times more potent than prednisolone (89). The above conversions

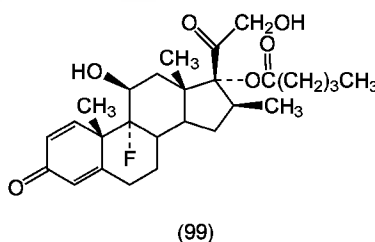


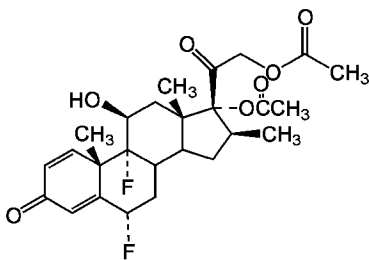
were applied to the betamethasone series, and the resultant β -methasone 17-valerate (99) had high topical activity in the McKenzie-Stoughton vasoconstrictor assay (90). The related β -methasone benzoate was developed independently (91). A number of ortho esters and 17,21-diester derived from $6\alpha,9\alpha$ -difluoroprednisolone were prepared (92) and were at least 1000 times more active than cortisol in the granuloma pouch assay. In addition, di-esters, eg, the 17-propionate 21-butyrate and 17-butyrate-21-acetate, were 3.5 times more potent than betamethasone valerate in vasoconstriction tests in humans. The 17-propionate and 17-butyrate of $6\alpha,9\alpha$ -difluoro-21-desoxyprednisolone were



2.5–3.0 times more potent than betamethasone valerate in the vasoconstriction assay in humans (93). In addition to the ortho ester route to 17-acylated corticoids, the facile and nearly quantitative diacylation of the corticoid side chain using a lower acid anhydride admixed with ortho phosphoric acid also was reported (94).

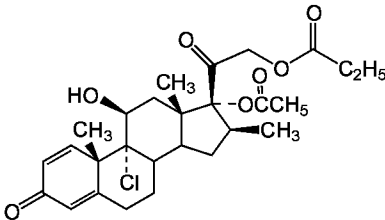
Recently developed 17-acylated corticoids include $6\alpha,9\alpha$ -difluoro-16 β -methyl-11 β 17 α ,21-trihydroxypregna-1,4-diene-3,20-dione diacetate (diflorasone diacetate) (100) (95); 9α -chloro-11 β ,17 α ,21-trihydroxy-16 β -methylpregna-1,4-diene-3,20-dione 17,21-dipropionate (beclomethasone dipropionate) (101) (96); and 21-chloro-11 β ,17 α -dihydroxy-9 α -fluoro-16 β -methylpregna-1,4-diene-3,20-dione 17-propionate (clobetasol propionate) (102) (97). In addition to the preceding corticoids which contain the dihydroxy acetone side chain, a number of products have been developed that lack an 17 α -hydroxyl moiety. 6α -Fluoro-16 α -methyl-1,4-pregnadiene-11 β ,21-diol-3,20-dione (flucortolone) (103) was the first product of this type to be introduced into therapy as an antiinflammatory agent (98). 11 β -Hydroxy-16 α ,17 α ,21-trimethylpregna-1,4-diene-3,20-dione (104) was compared to betamethasone valerate (99) in a number of antiinflammatory assays and they were approximately equipotent. In contrast to betamethasone valerate (99), however, it did not suppress adrenal function on systemic administration or reduce skin weight on topical administration.



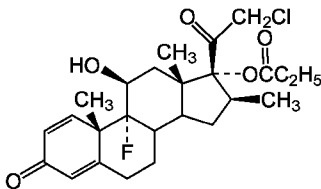


(100)

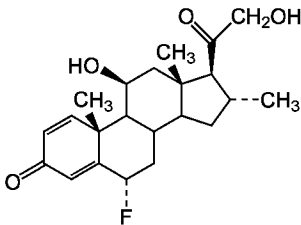
20-Ketopregnan-21-oic Acids, the 17β-Carboxy Androstanes, and the D-Homocorticoids. In the course of studies on the metabolism of fluocortolone (103) the formation of the water-soluble carboxylic acid (105, R = H) was reported. As a free 21-hydroxyl is not necessary for antiinflammatory activity, it was concluded that the esters (105, R = alkyl) of the preceding metabolite would possess antiinflammatory activity on topical administration but would be devoid of systemic activity when hydrolysis to the free acid occurs followed by



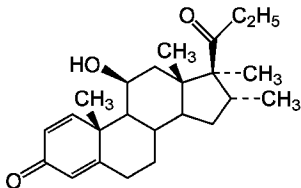
(101)



(102)



(103)



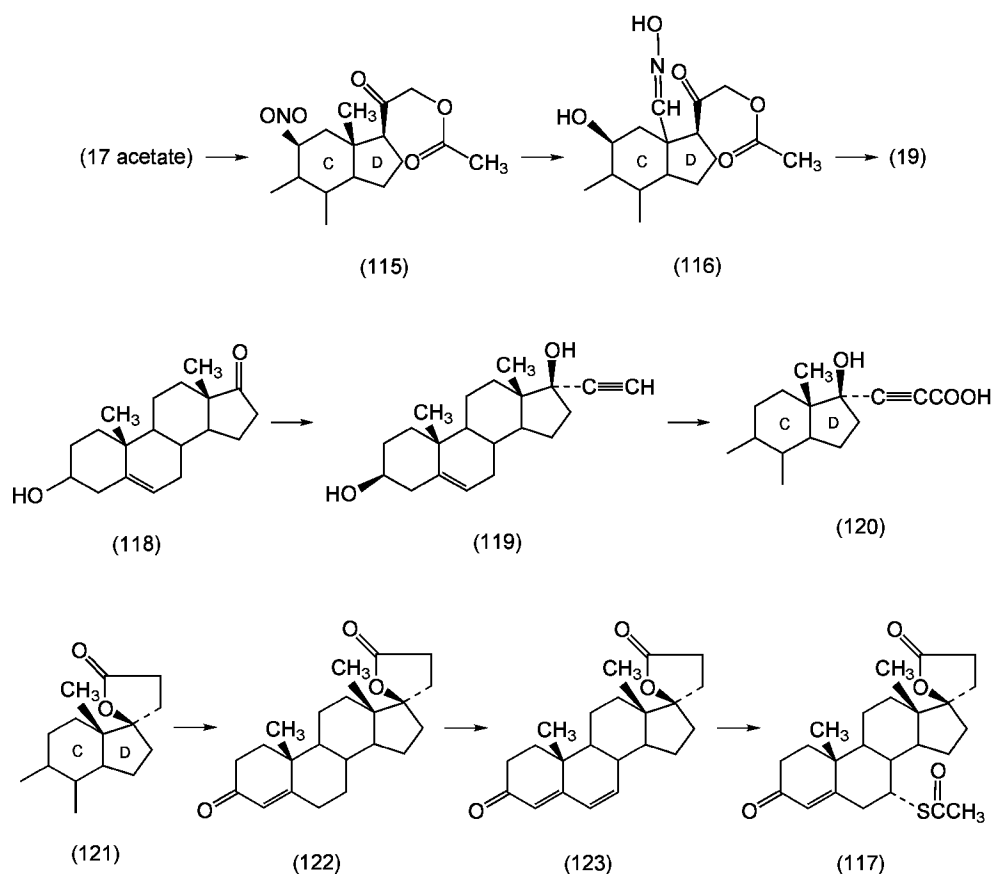
(104)

rapid elimination (100). The required esters (105, R = alkyl) were readily prepared by oxidation of fluocortolone (103) with alcoholic cupric acetate (101) whereby a mixture of the 20α- and 20β-hydroxypregnanoic esters (106) was obtained followed by oxidation with manganese dioxide to give the esters (105). Biological studies reveal that the esters (105) are highly active in the (topical) vasoconstriction assay but are devoid of systemic antiinflammatory activity as shown by the liver glycogen, paw edema, thymus involution, and other assays, and they were without mineralocorticoid activity. The butyl ester (fluocortin butyl ester) (105, R = C₄H₉) was selected for clinical study as a topical corticosteroid devoid of systemic activity (102). Urinary metabolites of prednisolone likewise contain products with β-glycolic and β-glyoxalic side chains. The methyl ester (106) of the latter in the paw edema assay is least effective if administered orally or intraperitoneally and most effective if administered topically or subcutaneously (103).

Another series of antiinflammatory carboxylic acids that are derived from cortienic acid (107), a minor adrenal metabolite, has been described (104,105). Esterification of both the 17α-hydroxyl group and the carboxylic acid of (107) were required to develop a compound of high topical potency with low systemic activity. Peak activity was generally associated with a 17α-propionoxy group and a 17β-fluoromethoxy carbonyl (eg, (108)), or 17β-methoxycarbonyl residue.

The D-homo-corticoids represent the last group of antiinflammatory steroids worthy of mention (106).

In 1973 D-homo corticosteroids (109–112), eg, D-homo-9α-fluoroprednisolone acetate (111) were reported to have antiinflammatory activity (107). Compounds such as 21-acetoxy-11β-fluoro-9α-chloro-17α-hydroxy-D-homo-pregn-4-en-3,20-dione (110) had especially strong topical activity with weak systemic activity (108). Other preparations of D-homocorticoids included



is synthesized from dehydroepiandrosterone (DHA) (118) (116). Ethynylation of DHA gives the 17 β -hydroxy-17 α -ethynyl derivative (119), which is converted using its ethynyl Grignard derivative and treatment with CO₂ into the 21-carboxylate (120). Hydrogenation leads to formation of the lactone (121), which is converted by the Oppenauer method into the 3-oxo- Δ^4 -lactone (122). The last compound is treated with chloranil in *t*-butanol to give the 4,6-diene (123) which adds thioacetic acid to give spironolactone (117). The market for mineralocorticoids is extremely modest as their main value lies in the treatment of Addison's disease.

Adrenal-Cortical Steroid Antagonists

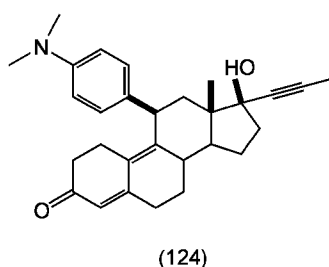
Antagonists of glucocorticoid and mineralocorticoid activity have found increased use clinically in the treatment of hypertension, Cushing's disease, and heightened intraocular pressure.

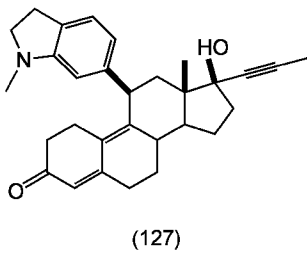
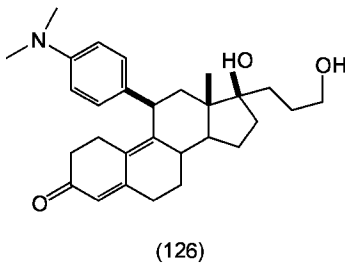
Some general structural features that lead to both observed differences in steroidal conformation (via x-ray structures) and to noticeable effects in bioactivity (RBAs, agonism vs antagonism, magnitude of activity, etc) are the following: (1) Unsaturation of the A-ring: Alteration of the A ring conformation. Except for the A-ring diazoles, all the GR antagonist known have the 4-ene-3-one structure. As this is shared with the most potent glucocorticoids, it is thought that this structural feature increases binding affinity to the steroid receptor (117). (2) The hydroxy groups at C-11, C-17, and C-21, although not necessary for a compound to bind to the GR, seem to be individually or collectively responsible for agonistic activity (though some compounds with hydroxyls at one or more of these positions (eg, RU-486 (124) and dexamethasone oxetanone (125), still act as GR antagonist). The absence of any of these hydroxyls may be one factor related to antagonistic activity (117), though not necessarily a determining factor. The lack of a hydroxyl at C-11 is a noticeable factor in the case of some GR antagonists. Where as all antiinflammatory glucocorticoid agonists have a hydroxyl at C-11 (or are readily solvolyzed to a C-11 OH), many antagonists though not all have no hydroxy moiety at C-11. For example, compounds that have shown antagonistic, eg, ZK 98299 (126), or mixed agonist activities, eg, RU-486 (124), each have substantially bulky groups at C-11.

An interesting comparison between agonist and antagonist to the GR is that of dexamethasone (74) (agonist) with dexamethasone oxetanone (125) (antagonist). The x-ray structures of these compounds indicate that all four rings overlap very well and thus should fit into the ligand binding site in a similar manner. As with differences in the magnitude of agonist activity, the agonist-antagonist distinction may be the result of chemical factors, including the hydrogen bond-donating or accepting ability of the ligand. In this example, both compounds can accept a hydrogen bond at C-20, though only the agonist is able to donate hydrogen bonds (117).

The most useful GR antagonist clinically is RU-486 (118,119), which also shows significant PR antagonism. This compound has been characterized more completely as a partial antagonist in certain cell lines (120,121), and has thus sometimes been called a type II antagonist. Although type II antagonist-occupied receptor can bind to DNA, it cannot interact productively with the DNA in a manner equivalent to interaction of an agonist-occupied receptor (122). RU-486 blocks the transactivation of the GR and inhibits the production of associated proteins, notably dexamethasone-induced extracellular LC-1 (55).

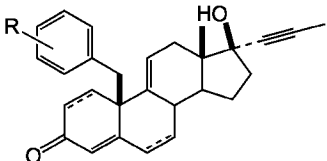
The defining feature of RU-486 is the 11 β -aryl moiety. Most steroidal GR antagonists include 11 β -aryl or other 11 β substituents. The tight binding affinity of RU-486 to the GR is presumed to be in part the result of a tight fit of this 11 β -aryl component in a lipophilic pocket within the GR hormone binding region. ZK 98299 is an 11 β -aryl GR antagonist very similar in structure to RU-486, though it is a pure (type I) antagonist. Like RU-486, it also has antiprogesterational activity. The 11 β -*N*-methylindol-5-yl analogue of RU-486 (127), another compound with GR antagonistic properties and a strong GR binding affinity, has a rotomer which fits well in the space (11 β -pocket) occupied by the 11 β -substituent of RU-486.





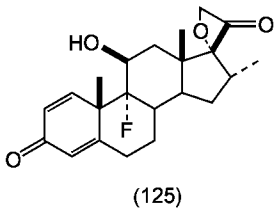
Another class of glucocorticoid antagonists, the 10β-androstanes, eg, RU-39305 (128) and RU-43044 (129), also bind tightly to the GR (Table 2). Relative binding affinities (RBA) to the GR (vs dexamethasone) do not lead to clear SARs in regard to unsaturation at Δ1 or Δ6. However, space-filling models indicate a similarity between the 3-D space occupied by the 10β- and 11β-compounds (123), although the aromatic moiety of RU-39305 (128) is not in the same orientation as that of RU-486. These compounds have very little affinity for the progesterone receptor (PR).

Table 2. Relative Binding Affinities of 10β-Androstanes^a

					
RU Code(Structure number)	Unsaturation	R	ThymusGR RBA ^b	PR RBA	ThymocytesIC ₅₀ nM
39305		H	57	<0.1	100
43044		<i>p</i> -CH ₃	130	<0.2	200
43065		<i>p</i> -CH ₃	17	<0.1	500
46759	Δ1, Δ6	<i>p</i> -CH ₃	33	<0.1	500
44068	Δ1	<i>p</i> -CH ₃	15	<0.1	>1000
44427	Δ6	<i>p</i> -CH ₃	3	<0.1	1000

^a Adapted from Ref. 123.
^b Dexamethasone — 100.

Research conducted by Simons using antiglucocorticoids, including compounds which covalently bind to the GR (124), eg, dexamethasone 21-mesylate, has better defined the structure and function of the GR. Spiro C-17 oxetanes have shown potent antiglucocorticoid activity in whole cell systems (125,126).



Although both the 10β- and the 11β-group have been found to be important in connection with many GR antagonists, neither is a requirement for antiglucocorticoid activity. Cortisolone (15) acts as a GR antagonist in the true sense of the term, blocking GR transactivation (127–129) although it does not include an 11β-substituent.

Spironolactone is the most clinically useful steroidal aldosterone antagonist, and unlike GR antagonists, this compound is utilized much more frequently than aldosterone agonists. Interfering with Na⁺ reabsorption and K⁺ secretion in the late distal segment, this compound is predominantly used with other diuretics. Canrenone, an olefinic metabolite of spironolactone, and potassium canrenoate, in which the C-17 lactone has been hydrolyzed open, are also potent mineralocorticoid antagonists.

BIBLIOGRAPHY

"Adrenal Cortical Hormones" under "Hormones" in *ECT* 1st ed., Vol. 7, pp. 495–513, by H. B. MacPhillamy, Ciba Pharmaceutical Products and T. F. Gallagher, Sloan-Kettering Institute for Cancer Research; "Steroids with Cortical Hormone Activity" in *ECT* 1st Suppl., pp. 849–888, by G. Anner and A. Wettstein, Ciba Limited, Basle; "Adrenal-Cortical Hormones" under "Hormones" in *ECT* 2nd ed., Vol. 11, pp. 77–93, by W. R. Eberlein, The Children's Hospital of Philadelphia; in *ECT* 3rd ed., Vol. 12, pp. 575–602, by V. Petrow, Consultant.

1. J. D. Baxter and G. G. Rousseau, in J. D. Baxter and G. G. Rousseau, ed., *Glucocorticoid Hormone Action: An Overview*, Vol., Springer-Verlag, New York, 1979, pp. 1–24.

2. P. S. Hench, E. C. Kendall, C. H. Slocumb, and H. F. Polley, *Proc. Staff Meet., Mayo Clinic* **24**, 181 (1949).

3. J. Fried and E. F. Sabo, *J. Am. Chem. Soc.* **75**, 2273 (1953).

4. J. Wepierre and J.-P. Marty, *Trends Pharmacol. Sci.*, 23 (1979).

5. M. E. Wolff, in Ref. 1, Vol. 12, pp. 97–107.

6. H. J. Lee, I. B. Taraporewala, and A. S. Heiman, *Med. Actual.* **25**, 577–588 (1989).

7. M. A. Avery and J. R. Woolfrey, in M. E. Wolfe, ed., *Burger's Medicinal Chemistry and Drug Discovery*, 5th ed., Vol. 5, John Wiley & Sons, Inc., New York, pp. 281–376.

8. H. P. Rang, M. M. Dale, P. Gardner, J. M. Ritter, *Pharmacology*, Churchill Livingstone, New York, 1995, pp. 444–447; B. G. Katzung, *Basic and Clinical Pharmacology*, Appleton & Lang, Norwalk, CT, 1995, p. 601.

9. J. R. Vane and R. M. Botting, *Inflam. Res.* **44**, 1–10 (1995).

10. P. J. Barnes and I. Adcock, *Trends Pharmacol. Sci.* **14**, 436–441 (1993).

11. G. K. McEvoy, ed., *AHFS Drug Information*, American Society of Health-System Pharmacists, Bethesda, Md., 1995, pp. 2439.

12. B. R. Olin, ed., *Drug Facts and Comparisons*, Facts and Comparisons, Inc., St. Louis, Mo., 1996, p. 122.

13. Ref. 11, p. 2439.

14. R. W. Brown, K. E. Chapman, P. Murad, C. R. Edwards, and J. R. Seckl, *Biochem. J.* **313**, 997–1005 (1996).

15. H. Vierhapper, K. Derfler, P. Nowotney, U. Hollenstein, and W. Waldhausl, *Acta Endocrinol. Copenh.* **125**, 160–164 (1991).

16. N. M. Drayer and C. J. P. Giroud, *Steroids* **5**, 289 (1965).

17. E. Gerhards, B. Nieuweboer, G. Schulz, and H. Gibian, *Acta Endocrinol.* **68**, 98 (1971).

18. H. L. Bradlow, B. Zumoff, C. Monder, H. J. Lee, and L. Hellman, *J. Clin. Endocrinol. Metab.* **37**, 811 (1973).

19. S. Edsbacker, P. Andersson, C. Lindberg, J. Paulson, A. Ryrfeldt, and A. Thalen, in D. J. Benford, J. W. Bridges, and G. G. Gibson, eds., *Metabolic Acetal Splitting of Budesonide: A Novel Inactivation Pathway for Topical Steroids*, Taylor and Francis, London, 1987, pp. 651–655.

20. G. Jonsson, A. Astrom, and P. Andersson, *Drug Metab. Disp.* **23**, 137–142 (1995).

21. G. M. Tomkins and P. J. Michael, *J. Biol. Chem.* **225**, 13 (1957).

22. E. M. Glenn, R. O. Stafford, S. C. Lyster, and B. J. Bowman, *Endocrinology* **61**, 128 (1957).

23. J. R. Florini, L. L. Smith, and D. A. Buyske, *J. Biol. Chem.* **236**, 1038 (1961).

24. R. M. Evans, *Science* **242**, 889–895 (1988).

25. J. Carlstedt-Duke, P.-E. Stromstedt, O. Wrange, T. Bergman, J.-A. Gustafsson, and H. Jornvall, *Proc. Natl. Acad. Sci. USA* **84**, 4437–4440 (1987).

26. M. Denis, J.-A. Gustafsson, and A.-C. Wikstrom, *J. Biol. Chem.* **263**, 18520–18523 (1988).

27. P. V. Bodine, E. S. Alnemri, and G. Litwack, *Receptor* **5**, 117–122 (1995).

28. L. F. Stancato, A. M. Silverstein, C. Gitler, B. Groner, and W. B. Pratt, *J. Biol. Chem.* **271**, 8831–8836 (1996).

29. C. Yu, N. Warriar, and M. V. Govindan, *Biochemistry* **34**, 14163–14173 (1995).

30. M. Schena, L. P. Freedman, and K. R. Yamamoto, *Genes Dev.* **3**, 1590–1601 (1989).

31. L. P. Freedman and co-workers, *Nature* **334**, 543–546 (1988).

32. J. Miller, A. D. McLachlan, and A. Klug, *EMBO J.* **4**, 1609–1614 (1985).

33. R. S. Brown, C. Sander, and P. Argos, *FEBS Lett.* **186**, 274 (1985).

34. B. F. Luisi, W. X. Xu, Z. Otwinowski, L. P. Freedman, K. R. Yamamoto, and P. B. Sigler, *Nature* **352**, 497–505 (1991).

35. M. A. Eriksson and L. Nilsson, *J. Mol. Biol.* **253**, 453–472 (1995).

36. T. Hard and co-workers, *Science* **249**, 157–160 (1990).

37. K. Dahlman-Wright and co-workers, *Proc. Natl. Acad. Sci. USA* **92**, 1699–1703 (1995).

38. I. J. McEwan, K. Dalman-Wright, T. Amlof, J. Ford, A. P. Wright, and J. A. Gustafsson, *Mutat. Res.* **333**, 15–22 (1995).

39. B. F. O'Malley, in T. Tsi, ed., *Glucocorticoid Action*, Vol. , Academic Press, New York, 1995, pp. 455–492.

40. M. Brink, B. M. Humbel, E. R. DeKloet, and R. Vandriel, *Endocrinology* **130**, 3575–3581 (1992).

41. G. Akner, A.-C. Wikstroem, and J.-A. Gustafsson, *J. Steroid Biochem. Mol. Biol.* **52**, 1–16 (1995).

42. C. M. Jewell and co-workers, *J. Steroid Biochem. Mol. Biol.* **55**, 135–146 (1995).

43. D. Picard and K. R. Yamamoto, *EMBO J.* **6**, 3333–3340 (1987).

44. K. A. Hutchison, K. D. Dittmar, and W. B. Pratt, *J. Biol. Chem.* **269**, 27894–27899 (1994).

45. K. A. Hutchinson, L. F. Sanncato, J. K. Owens-Grillo, J. L. Johnson, P. Krishna, D. O. Toft, and W. B. Pratt, *J. Biol. Chem.*, **270**, 18841–18847 (1995).

46. E. H. Bresnick, F. C. Dalman, E. R. Sanchez, and W. B. Pratt, *J. Biol. Chem.* **264**, 4992–4997 (1989).

47. C. C. Landel, P. J. Kuchner, and G. L. Greene, *Env. Health Perspect.*, **103**, 23–28 (1995).

48. J. Drouin and co-workers, *Mol. Endocrinol.* **6**, 1299–1309 (1992).

49. R. J. Flower and N. J. Rothwell, *Trends Pharmacol. Sci.* **15**, 71–76 (1994).

50. B. P. Wallner and co-workers, *Nature* **320**, 7 (1986).

51. J. D. McLeod, A. Goodall, P. Jelic, and C. Bolton, *Biochem. Pharmacol.* **50**, 1103–1107 (1995).

52. S. F. Smith, T. D. Tetley, A. K. Datta, T. Smithe, A. Guz, and R. J. Flower, *J. Appl. Physiol.* **79**, 121–128 (1995).

53. J. D. Croxtall, Q. Choudhury, H. Tokumoto, and R. J. Flower, *Biochem. Pharmacol.* **50**, 465–474 (1995).

54. M. Perretti, S. K. Wheller, Q. Choudhury, J. D. Croxtall, and R. J. Flower, *Biochem. Pharmacol.* **50**, 1037–1042 (1995).

55. A. Ahluwalia, R. W. Mohamed, and R. J. Flower, *Biochem. Pharmacol.* **48**, 1647–1654 (1994).

56. G. Cirino, R. J. Flower, J. L. Browning, L. K. Sinclair, and R. B. Pepinsky, *Nature* **329**, 270–272 (1987).

57. G. Cirino, S. H. Peers, R. J. Flower, J. L. Browning, and R. B. Pepinski, *Proc. Natl. Acad. Sci. USA* **86**, 3428–3432 (1989).

58. P. J. Barnes, *Biochem. Soc. Trans.* **23**, 940–945 (1995).

59. G. S. Duncan, S. H. Peers, F. Carey, R. Forder, and R. J. Flower, *Br. J. Pharmacol.* **108**, 62–65 (1993).

60. A. Ahluwalia, P. Newbold, S. D. Brain, and R. J. Flower, *Eur. J. Pharmacol.* **283**, 193–198 (1995).

61. A. D. Taylor, H. D. Loxley, R. J. Flower, and J. C. Buckingham, *Neuroendocrinology* **62**, 19–31 (1995).

62. H. Tanaka, *Shindan to Chiryō* **83**, 1205–1209 (1995).

63. G. E. Hill, S. Snider, T. A. Galbraith, S. Forst, and R. A. Robbins, *Am. J. Respir. Crit. Care. Med.* **152**, 1791–1795 (1995).

64. G. Michel, K. Nowok, A. Beetz, C. Ried, L. Kemeny, and T. Ruzicka, *Skin Pharmacol.* **8**, 215–220 (1995).

65. J. M. Cavaillon and C. R. Seances, *Soc. Biol. Fil.*, **189**, 531–544 (1995).

66. K. Kijima, H. Matsubaro, S. Murasowo, S. Muruyama, U. Mori, and M. Inada, *Biochem. Biophys. Res. Commun.*, **216**, 359–366 (1995); N. Applezweig, ed., *Steroid Drugs*, McGraw-Hill, Inc., New York, 1962.

67. D. H. Peterson, H. C. Murray, S. H. Eppstein, L. M. Reineke, A. Weintraub, P. D. Meister, and H. M. Leigh, *J. Am. Chem. Soc.* **74**, 5933 (1952); S. H. Eppstein, P. D. Meister, D. H. Peterson, H. C. Murray, H. M. Leigh, D. A. Lyttle, L. M. Reineke, and A. Weintraub, *J. Am. Chem. Soc.* **75**, 408 (1953).

68. H. L. Herzog, A. Nobile, S. Tolksdorf, W. Charney, E. B. Hershberg, and P. I. Pedman, *Science* **121**, 176 (1955); A Nobile, W. Charney, P. L. Perlman, H. L. Herzog, C. C. Paine, M. E. Tully, M. A. Jevnik, and E. B. Hershberg, *J. Am. Chem. Soc.* **77**, 4184 (1955).

69. F. W. Heyl and M. E. Herr, *J. Am. Chem. Soc.* **72**, 2617 (1950).

70. M. E. Herr and F. W. Heyl, *J. Am. Chem. Soc.* **74**, 3627 (1952).
71. J. A. Hogg, P. F. Beal, A. H. Nathan, F. H. Lincoln, W. P. Schneider, B. J. Magedein, A. R. Hanze, and R. W. Jackson, *J. Am. Chem. Soc.* **77**, 4436 (1955); J. A. Hogg, F. H. Lincoln, A. H. Nathan, A. R. Hanze, W. P. Schneider, P. F. Beal, and J. Korman, *J. Am. Chem. Soc.* **77**, 4438 (1955); Brit. Pat. 873,608 (July 26, 1961), (to The Upjohn Co.).
72. **Belg. Pat. 716,161** (June 12, 1968) (to The Upjohn Co.).
73. S. Bernstein, R. H. Lenhard, W. S. Allen, M. Heller, R. Littell, S. M. Stolar, L. I. Feldman, and R. H. Blank, *J. Am. Chem. Soc.* **78**, 5693 (1956).
74. J. Fried, A. Borman, W. B. Kessler, P. Grabovitch, and E. F. Sabo, *J. Am. Chem. Soc.* **80**, 2338 (1958).
75. I. Ringler, S. Maurer, and E. Heyder, *Proc. Soc. Exper. Biol. Med.* **107**, 451 (1961); E. Mascietti-Coriandoli and A. Fraia, *Arzneim. Forsch.* **20**, 111 (1970).
76. J. A. Hogg, F. H. Lincoln, R. W. Jackson, and W. P. Schneider, *J. Am. Chem. Soc.* **77**, 6401 (1955).
77. G. B. Spero, J. L. Thompson, B. J. Magedein, A. R. Hanze, H. C. Murray, O. K. Sebek, and J. A. Hogg, *J. Am. Chem. Soc.* **78**, 6213 (1956).
78. G. Bianchi, A. David, B. Ellis, V. Petrow, S. Waddington-Feather, and E. J. Woodward, *J. Pharm. Pharmacol.* **13**, 355 (1961).
79. **Brit. Pat. 864,381** (Mar. 29–Apr. 6, 1961), F. H. Lincoln, Jr., W. P. Schneider, and G. B. Spero (to The Upjohn Co.).
80. G. Arth, D. Johnston, J. Fried, W. Spooncer, D. Hoff, and L. Sarrett, *J. Am. Chem. Soc.* **80**, 3160 (1958); E. Oliveto, R. Rausser, A. Nussbaum, W. Gebert, E. B. Hershberg, S. Tolksdorf, M. Eisler, P. Perlman, and M. Pechet, *J. Am. Chem. Soc.* **80**, 4428 (1958); D. Taub, R. Hoffsommer, H. Slates, and N. Wendler, *J. Am. Chem. Soc.* **80**, 4335 (1958); E. Oliveto, *N.Y. Acad. Sci.* **82**, 809 (1959).
81. R. Marker and H. Crocks, *J. Am. Chem. Soc.* **64**, 1280 (1942).
82. E. Oliveto, R. Rausser, M. Eisler, P. Perlman, and M. Pechet, *J. Am. Chem. Soc.* **80**, 4431 (1958).
83. A. Wettstein, *Helv. Chim. Acta* **27**, 1803 (1944).
84. T. F. Gallagher and W. P. Long, *J. Biol. Chem.* **162**, 495, 521 (1946); J. Elks, G. H. Phillips, T. Walker, and L. J. Wyman, *J. Chem. Soc.*, 4330 (1956); J. H. Chapman, J. Elks, G. H. Phillips, and L. J. Wyman, *J. Chem. Soc.*, 4344 (1956).
85. **Ger. Pat. 1,263,765** (June 21, 1968), K. Immscher and D. Orth (to E. Merck A.G.).
86. J. A. Edwards, H. J. Ringold, and C. Djerassi, *J. Am. Chem. Soc.* **82**, 2318 (1960).
87. J. S. Mills, A. Bowers, C. Djerassi, and H. J. Ringold, *J. Am. Chem. Soc.* **82**, 3399 (1960).
88. J. S. Mills, A. Bowers, C. C. Campillo, C. Djerassi, and H. J. Ringold, *J. Am. Chem. Soc.* **81**, 1264 (1959).
89. R. Garci, R. Vitali, and A. Ercoli, *Tetrahedron Lett.* (13), 448 (1961); **U.S. Pat. 3,147,249** (Sept. 1, 1964), A. Ercoli and R. Gardi (to Francesco Vismara S.p.A.); R. Gardi, R. Vitali, and A. Ercoli, *Gazz. Chim. Ital.* **93**, 413, 431 (1963).
90. **U.S. Pat. 3,312,590** (Apr. 4, 1967), J. Elks, P. J. May, and N. G. Weir (to Glaxo Laboratories Ltd.).
91. **U.S. Pat. 3,529,060** (Sept. 15, 1970), A. Ercoli, R. Gardi, and C. Brianza (to Warner Lambert Pharm. Co.).
92. R. Gardi, R. Vitali, G. Falconi, and A. Ercoli, *J. Med. Chem.* **15**, 556 (1972).
93. R. Vitali, S. Gladiali, G. Falconi, G. Celasco, and R. Gardi, *J. Med. Chem.* **20**, 853 (1977).
94. **U.S. Pat. 4,154,748** (May 15, 1979), V. H. Van Rhee and J. B. Heather (to The Upjohn Company).
95. R. S. P. Hsi and D. E. Ayer, *J. Labelled Compd. Radiopharm.* **14**, 99 (1978).
96. **Ger. Pat. 1,904,324** (Sept. 11, 1969), G. H. Phillips (to Glaxo Laboratories Ltd.).
97. **Ger. Pat. 1,902,340** (Sept. 11, 1969), J. Elks and G. H. Phillips (to Glaxo Laboratories Ltd.).
98. A. Domíñico, H. Gibian, U. Kerb, K. Kieslich, M. Kramer, F. Neumann, and G. Raspñ, *Arzneim. Forsch.* **15**, 46 (1965).
99. K. H. Burdick, *Arch. Derm.* **98**, 684 (1968).
100. H. Laurent, E. Gerhards, and R. Wiechert, *J. Steroid Biochem.* **6**, 185 (1975); **U.S. Pat. 3,824,260** (July 16, 1974), H. Laurent, R. Wiechert, K. Prezewowsky, H. Hofmeister, E. Gerhards, K. H. Kolb, and K. Mengel (to Schering A.G.); **U.S. Pat. 3,875,194** (Apr. 1, 1975), H. Laurent and R. Wiechert (to Schering A.G.).
101. M. L. Lewbart and V. R. Mattox, *J. Org. Chem.* **28**, 1779 (1963).
102. H. Laurent, E. Gerhards, and R. Wiechert, *Arzneim. Forsch. II* **27**(11a), 2187 (1977); J. F. Kapp, H. Koch, M. Tolpert, H. J. Kessler, and E. Gerhards, *Arzneim. Forsch.* **27**, 2191; H. J. Kessler, D. Haude, C. Okon, and P. Schumann, *Arzneim. Forsch.* **27**, 2214; P. Guenzel, *Arzneim. Forsch.* **27**, 2217.
103. H. J. Lee, R. W. Trotter, J. McMillan, and P. Barnes, *Pharmacologist* **18**, 218 (1976).
104. B. M. Bain, P. J. May, G. H. Phillips, and E. A. Wollett, *J. Steroid Biochem.* **5**, 299 (1974); **Ger. Pat. 2,202,691** (Aug. 3, 1972), G. H. Phillips and P. J. May (to Glaxo Laboratories Ltd.); **U.S. Pat. 3,856,828** (Dec. 24, 1974), G. H. Phillips, P. J. May, B. McDonald, and E. A. Woollett (to Glaxo Laboratories Ltd.).
105. H. Levy, I. Cargill, C. H. Cha, B. Hood, and J. J. Carlo, *Steroids* **5**, 131 (1965).
106. E.-E. Baulieu, *J. Am. Med. Assoc.* **234**, 404 (1975); I. S. Edelman, *J. Steroid Biochem.* **6**, 147 (1975); K. Lbbke, E. Schellinger, and M. Tnpert, *Angew. Chem. Int. Ed. Eng.* **15**, 741 (1976); G. G. Rousseau, *J. Steroid Biochem.* **6**, 75 (1975).
107. **Neth. Pat. 7,304,193** (Oct. 2, 1973) (to Hoffmann-La Roche & Co.).
108. **U.S. Pat. 3,939,193** (Feb. 17, 1976), **U.S. Pat. 4,036,874** (July 19, 1977), and **U.S. Pat. 4,026,923** (May 31, 1977), L. Alig, A. Furst, and M. Muller (to Hoffmann-La Roche & Co.).
109. **U.S. Pat. 4,033,995** (July 5, 1977), R. C. Nickolson, U. Kerb, and R. Wiechert (to Schering A.G.).
110. L. Alig, A. Fuerst, P. Keller, M. Mueller, U. Kerb, and R. Wiechert, *J. Steroid Biochem.* **5**, 298 (1974).
111. **U.S. Pats. 4,026,918** (May 31, 1977), A. Furst and co-workers; **U.S. Pat. 4,029,692** (June 14, 1977), L. Alig and co-workers.
112. S. A. Simpson and J. F. Tait, *Endocrinology* **50**, 150 (1952); S. A. Simpson, J. F. Tait, A. Wettstein, R. Neher, J. von Euw, and T. Reichstein, *Experientia* **9**, 333 (1953); S. A. Simpson, J. F. Tait, A. Wettstein, R. Neher, J. von Euw, O. Schindler, and T. Reichstein, *Helv. Chim. Acta* **37**, 1163 (1954).
113. S. A. Simpson, J. F. Tait, A. Wettstein, R. Neher, J. von Euw, O. Schindler, and T. Reichstein, *Experientia* **10**, 132 (1954); *Helv. Chim. Acta* **37**, 1200 (1954).
114. D. H. R. Barton, J. M. Beaton, L. E. Geller, and M. M. Pechet, *J. Am. Chem. Soc.* **82**, 2640 (1960); D. H. R. Barton and J. M. Beato, *J. Am. Chem. Soc.* **82**, 2641 (1960).
115. A. Y. Vagnucci and A. P. Shapiro, *Metabolism* **23**, 273 (1974); *J. Endocrinol.* **81**, 1P (1979).
116. J. A. Cella and C. M. Kagawa, *J. Am. Chem. Soc.* **79**, 4808 (1957).
117. W. L. Duax, J. F. Griffin, C. M. Weeks, and Z. Wawrzak, *J. Steroid Biochem.* **31**, 481–492 (1988).
118. D. Philibert, T. Deraedt, and G. Teutsch, eds., *RU-38486: A Potent Antiglucocorticoid In Vivo*, Tokyo, 1981, pp.
119. D. Gagne, M. Pons, and D. Philibert, *J. Steroid Biochem.* **23**, 247–251 (1985).
120. S. K. Nordeen, B. J. Bona, C. A. Beck, D. P. Edwards, K. C. Borror, and D. B. DeFranco, *Steroids* **60**, 97–104 (1995).
121. S. Zhang and M. Danielsen, *Recent Prog. Horm. Res.* **50**, 429 (1995).
122. J. S. Mymryk and T. K. Archer, *Mol. Endocrinol.* **9**, 1825–1834 (1995).
123. D. Philibert, G. Costerousse, L. Gaillard-Mogulewsky, L. Nedelec, F. Nique, C. Toumemine, and G. Teutsch, in M. K. Agarwal, ed., *From RU 38486 Towards Dissociated Antiglucocorticoid and Antiprogesterone*, Vol. 19, Karger, Basel, Switzerland, 1991, pp. 1–17.
124. S. S. Simons and E. B. Thompson, *Proc. Natl. Acad. Sci. USA*, **78**, 3541–3545 (1981).
125. M. Pons and S. S. Simons, Jr., *J. Org. Chem.* **46**, 3262–3264 (1981).
126. S. S. Simons, Jr., E. B. Thompson and D. F. Johnson, *Proc. Natl. Acad. Sci. USA* **77**, 5167–5171 (1980).
127. R. W. Turnell, N. Kaiser, R. J. Millholland, and F. Rosen, *J. Biol. Chem.* **249**, 1133–1138 (1974).

128. C. R. Wira and A. Munck, *J. Biol. Chem.* **249**, 5328–5336 (1974).
129. W. L. Duax and J. F. Griffin, in B. Testa, ed., *The Structure and Receptor Binding of Steroid Hormones*, Vol. 18, Academic Press, London, 1989, pp. 116–132.

Mitchell A. Avery
John R. Woolfrey
University of Mississippi

LITHOGRAPHIC RESISTS

Lithography in its original meaning describes a method of printing where a master pattern or design first is formed on a stone or metal surface using a hydrophobic material, the pattern is wetted with a hydrophobic ink, and finally the inked pattern is transferred to paper to produce copies of the master. Today, the meaning of the term has broadened to include a number of methods for replicating a predetermined pattern on a substrate, for an increasing variety of purposes. In most cases, replication is effected by first coating the substrate with a photosensitive polymer film termed a *lithographic resist*, and then exposing the film to a pattern of light (1–3). Photochemically induced changes in areas of the film exposed to light alter solubility to such extent that either exposed or unexposed areas can be selectively dissolved to reveal portions of the substrate surface. If the exposed areas of resist become insoluble, then the resist is termed negative-acting; if solubility of the exposed regions increases then the resist is termed positive-acting.

In certain applications, the resist film becomes a permanent, functional component of the device being constructed, typically serving as an electrical insulator or protective encapsulant. In such cases the patterned film often is treated to enhance chemical resistance and mechanical properties (4). Most commonly, however, the resist pattern is a temporary template, and is removed or stripped after the image has been transferred to the substrate. Numerous methods of image transfer have been practiced, including plating and deposition (*additive processing*), chemical etching or physical milling to erode the exposed substrate (*subtractive processing*) or bombardment with energetic ions which become implanted in the substrate surface. Figure 1 summarizes the steps comprising the lithographic process.

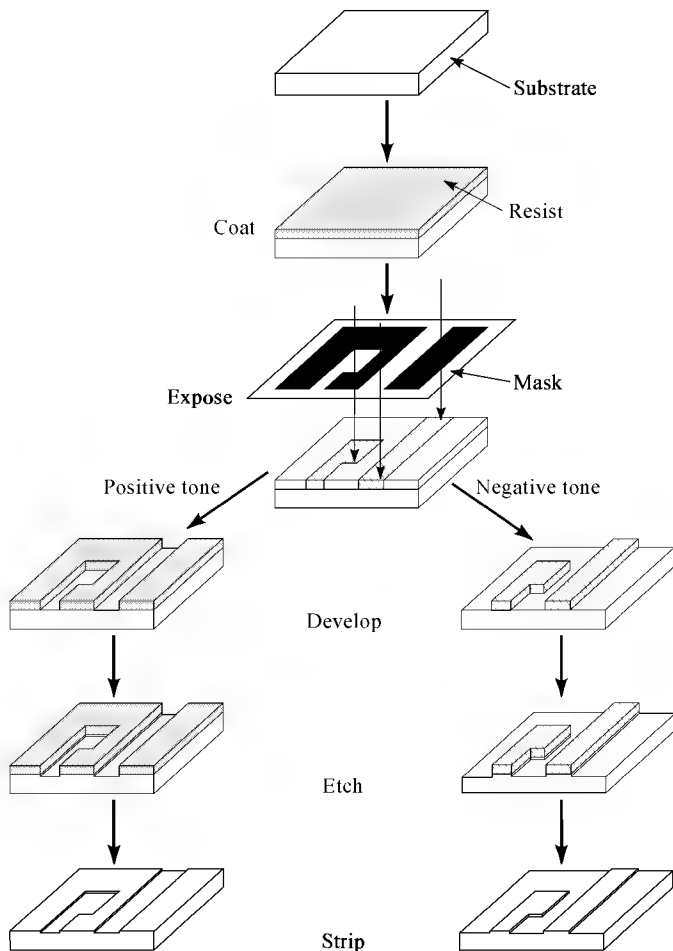


Fig. 1. The lithographic process. A substrate is coated with a photosensitive polymer film called a resist. A mask with transparent and opaque areas directs radiation to preselected regions of the resist film. Depending on resist characteristics, exposed or unexposed portions of the film are removed using a developer solvent. The resulting pattern is then transferred to the substrate surface and the resist is stripped.

Lithographic resists enable a host of technological advances that have had far-ranging impacts. For example, the rapid progress in microelectronics technology in large part stems from refinement of the lithographic techniques used to fabricate computer integrated circuits. The ability to pack ever-increasing numbers of discrete electronic components in a given area of an integrated circuit, when coupled with the economies of scale that result from high volume manufacturing, has led to both improved performance and lower cost with each succeeding generation (5). Today the annual global market for semiconductors far exceeds \$100 billion U.S.

The advantages of miniaturization are now being exploited in areas beyond microelectronics. Adaptation of materials and processes originally devised for semiconductor manufacture has allowed fabrication of sensors (for example, pressure meters and accelerometers used in the automotive industry) (6,7), complex optical (8) and micromechanical (6,7,9) assemblies, and devices for medical diagnostics (6,7,10) using lithographic resists.

Essential Attributes of Lithographic Resists

Regardless of the specific application, all resists must display certain fundamental functional properties:

- (1) The resist composition must form *uniform, defect-free films* on the substrate of interest. To achieve this, the resist components must be soluble in and inert toward the selected coating solvent or medium, the formulation must form glassy films without component precipitation, separation or dewetting, and must be stable during post-application heating steps. A variety of different coating techniques have been applied, including

- spray-coating, curtain-coating, spin-coating or whirling, dip-coating, dry-film lamination, and most recently, electrodeposition onto a conducting substrate from an aqueous emulsion of resist components.
- (2) The coated film must display adequate *adhesion* to the substrate after coating, and through the develop and image transfer steps. Adherence is influenced by resist composition, substrate surface composition, preparation and cleanliness, coating conditions, develop and etch process details, and, with negative cross-linking resists, by the exposure dose. Adhesion loss in mild cases is manifested by the lifting of a resist pattern at its perimeter (often visible under microscopic examination as a contrasting band) and in extreme cases by complete detachment of the resist pattern.
 - (3) The resist must have suitable *radiation sensitivity*. Today's exposure tools are so costly that tool throughput is a key measure of performance. The overall time to expose a resist film is the sum of the times to load and position the substrate in the exposure tool, to align the substrate and the mask, to irradiate the film, and to unload the complete part. In the optimum case the resist exhibits sufficient radiation sensitivity so that the fraction of the overall cycle apportioned to irradiate the film does not limit the number of substrates exposed in a given period of time.
 - (4) The resist must provide *high fidelity reproduction* of the mask image on the substrate. This attribute is generally quantified as the resist's resolution, that is, the smallest feature that can be consistently printed, and its contrast, which is a fundamental measure of how the resist response changes with increasing radiation dose. In high resolution lithography, diffraction and imperfections in the optical system cause stray radiation to be scattered into regions of the resist pattern that nominally are unexposed. A resist with high contrast can effectively reject the stray radiation and yield a more accurate replica of the original mask. Resist contrast and resolution are influenced by its composition, by process conditions, and exposure tool characteristics.
 - (5) The patterned resist must provide an effective *protective barrier* during image transfer. The image transfer step frequently includes harsh chemical or physical treatments that erode or otherwise degrade the resist pattern. A resist's capacity to withstand the more vigorous image transfer processes (eg, plasma etching) can vary widely with its composition.
 - (6) After image transfer, the patterned resist must be readily and completely *removable* without substrate damage. The pattern often can be stripped from the substrate with a mild organic solvent. Proprietary stripper formulations or plasma oxidation treatments are utilized when the imaging chemistry or image transfer process has insolubilized the pattern.

Historical Development of Resist Materials

Most modern lithographic resists are evolved from materials first developed for the printing industry (11). Increasingly specialized resists and processing techniques have been introduced as applications of lithographic technology have grown in scope and sophistication. Figure 2a presents a time line noting significant events in this evolution; many of these are described in detail in this section. Figure 2b shows one representation of Moore's law (5), which traces the evolution of semiconductor lithographic technology, as characterized by the minimum feature size versus the year of commercial introduction of the corresponding device generation.

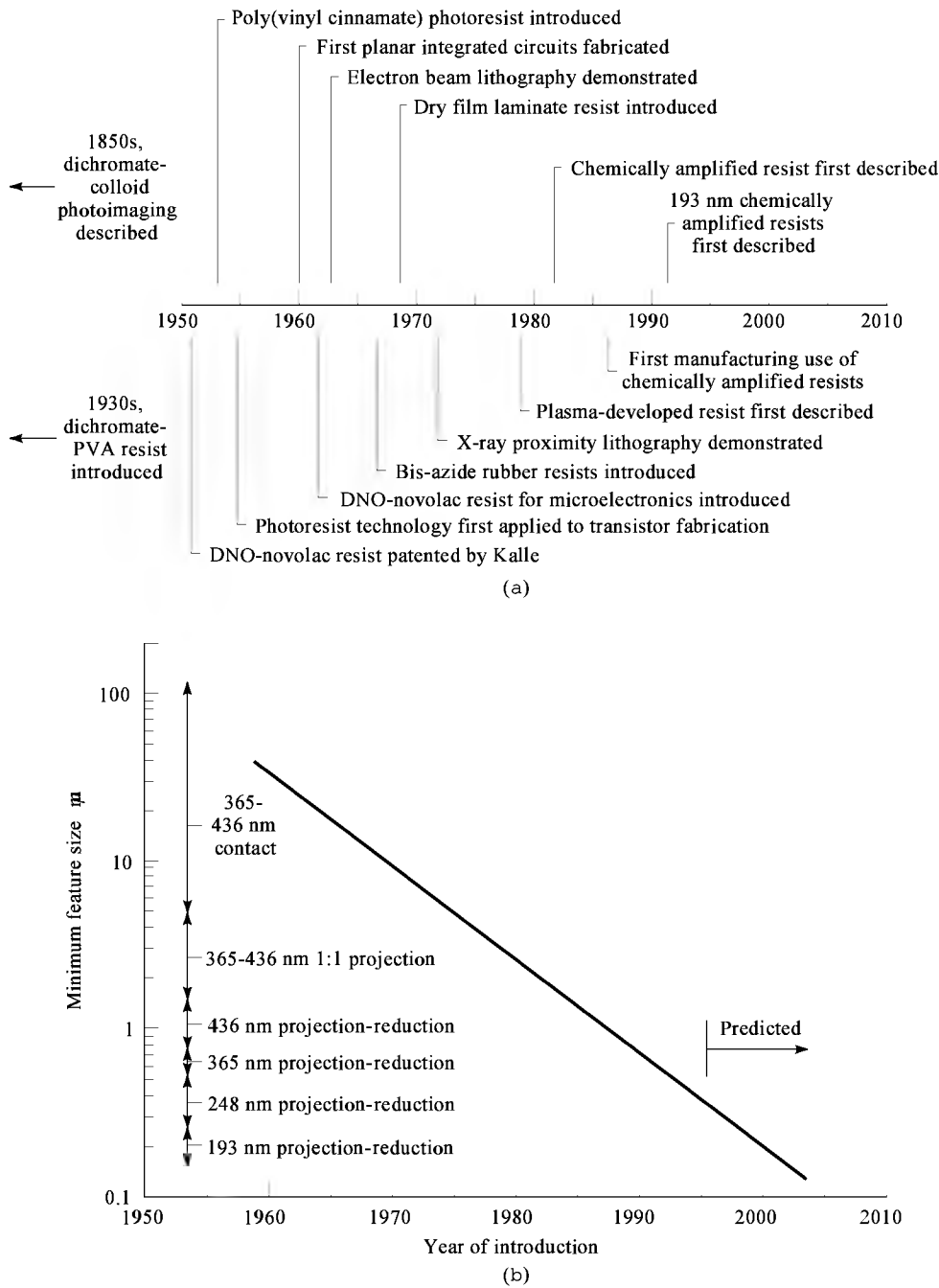


Fig. 2. Evolution of lithographic technology. (a) A time line noting significant events in the evolution of lithographic technology. (b) The minimum feature

size used in a commercial microelectronic device is plotted against its year of introduction. Also indicated are the exposure tool technologies used to produce the different device generators. In projection lithography, printed resist features can be either the same size as those on the mask (1:1), or can be a factor of 4–10 smaller (reduction).

Dichromated Resists. The first compositions widely used as photoresists combine a photosensitive dichromate salt (usually ammonium dichromate) with a water-soluble polymer of biologic origin such as gelatin, egg albumin (proteins), or gum arabic (a starch). Later, synthetic polymers such as poly(vinyl alcohol) also were used (11,12). Irradiation with uv light (λ in the range of 360–380 nm using, for example, a carbon arc lamp) leads to photoinitiated oxidation of the polymer and reduction of dichromate to Cr(III). The photoinduced chemistry renders exposed areas insoluble in aqueous developing solutions. The photochemical mechanism of dichromate sensitization of PVA (summarized in Fig. 3) has been studied in detail (13).

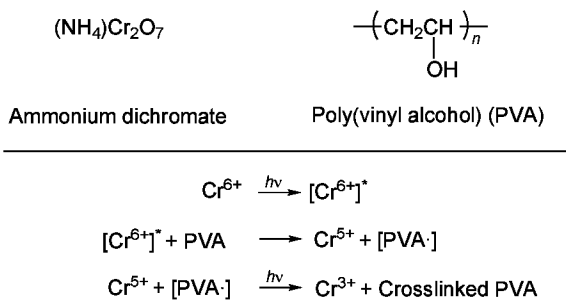
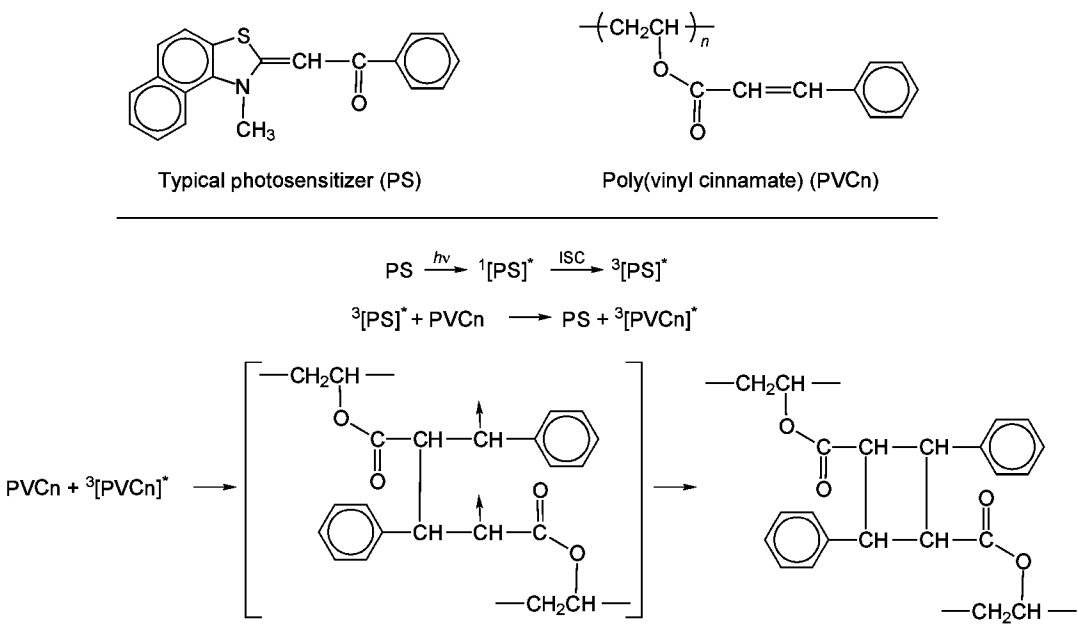


Fig. 3. Chemistry of dichromated poly(vinyl alcohol) resist. Initially the dichromate ion absorbs light; the light-activated Cr^{6+} species undergoes an electron-transfer reaction with the PVA matrix to form a polymer radical. These undergo further reactions to form Cr^{3+} and a cross-linked matrix (13).

Poly(vinyl cinnamate) Resists. Dichromated resists exhibit numerous shortcomings which include lot-to-lot variability of the components, aging of the formulated resists in solution and in coated form, poor process stability (due to a sensitivity to variations in temperature and humidity), and intrinsically low photosensitivity requiring long exposure times for adequate insolubilization.

In the 1950s resist systems with substantially improved processing characteristics were developed. The first commercially available member of this class, KPR, was introduced in 1953 by Eastman Kodak. Originally targeted for printing applications, KPR was used in the fabrication of circuit boards and semiconductor devices later in the decade. This material is a cross-linking system based on the photodimerization of poly(vinyl cinnamate) chains. Pendant carbon–carbon double bonds on adjacent polymer strands undergo photocyclization to form a cross-linked, insoluble network in exposed areas (Fig. 4). Although poly(vinyl cinnamate) is intrinsically photoactive, its electronic absorption peaks near a wavelength of 290 nm, and absorbs only weakly at wavelengths longer than 350 nm, where the lamps commonly used in exposure tooling (eg, a high pressure mercury arc) show strong emission. Its spectral sensitivity can be extended to longer wavelengths by addition of an appropriate absorbing dye. The dye is believed to act as a triplet photosensitizer, transferring excitation energy to a cinnamate acceptor. The intermediate triplet cinnamate adds to a ground state cinnamate to form a biradical species, which closes to produce the final cross-linked cyclic structure.



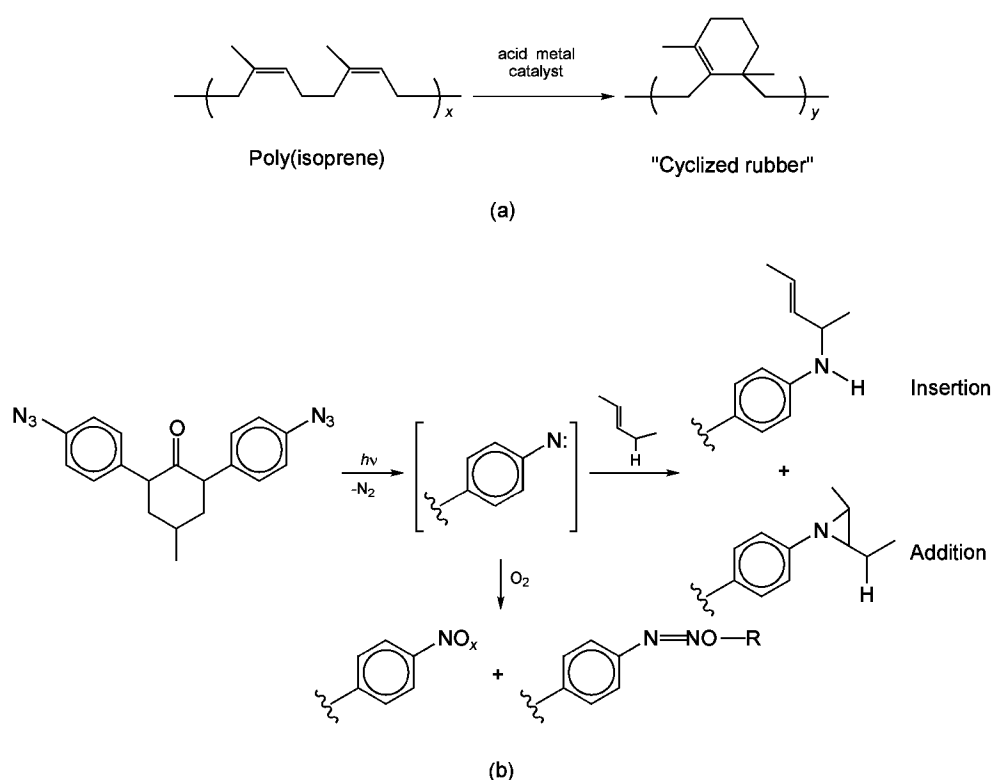


Fig. 5. Chemistry of cyclized rubber–bis-azide negative acting resist. (a) Preparation of cyclized rubber resin from polyisoprene; (b) photochemistry of aromatic bis-azide sensitizers. The primary photoproduct is a highly reactive nitrene which may combine with molecular oxygen to form oxygenated products, or may react with the resin matrix by addition or insertion to form polymer–polymer linkages.

During exposure of the formulated resist, azide groups in the sensitizer undergo sequential loss of molecular nitrogen. Each fragmentation produces a highly reactive nitrene with two unpaired electrons. The nitrene rapidly deactivates by a number of pathways (Fig. 5b). Cross-linking results when nitrene intermediates add to residual carbon–carbon double bonds or insert into carbon–hydrogen bonds, converting the matrix into an insoluble network.

First introduced for printing applications, bis-azide–rubber resists were soon adopted by the developing microelectronics industry and used extensively for device fabrication through the 1970s. However, as the semiconductor industry refined manufacturing processes to further miniaturize integrated circuits, the limitations of this chemistry became apparent.

First, molecular oxygen dissolved in the film degrades the imaging chemistry. The nitrene intermediate reacts readily with oxygen to form azoxy and nitroso products (Fig. 5b). This oxidation competes with addition and insertion, reducing the extent of polymer–polymer cross-linking. The adoption by the semiconductor industry of proximity- and projection-exposure methods (Fig. 6) in the 1970s magnified the impact of this oxygen effect. In the traditional contact technique the mask acts as a physical barrier preventing diffusion of atmospheric oxygen into the resist film during exposure, so that molecular oxygen that can react with the nitrene is limited to the small amount present at the moment of mask contact. Proximity and projection exposure techniques reduce device and mask damage by avoiding their direct mechanical contact. Atmospheric oxygen is then free to diffuse into the film, typically at a rate comparable with that of nitrene formation. The overall result is an increase in the effective concentration of oxygen in the film and accentuation of oxygen's inhibitory effect. Oxygen inhibition is manifested as excessive film thinning (due to suppression of cross-linking near the air–polymer interface) and increased dose requirements compared to contact exposure conditions.

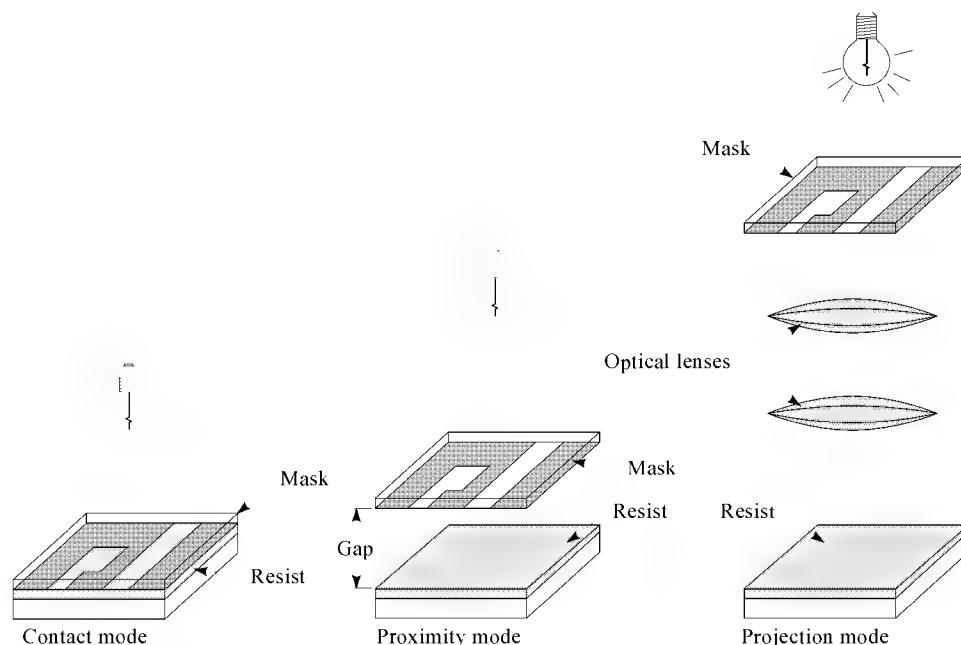


Fig. 6. Contact, proximity, and projection modes of optical lithography. Although contact mode generally produces the sharpest light image, damage to mask and substrate make this impractical for most electronic device applications. In proximity mode, a narrow, uniform gap between mask and substrate minimizes damage, but blurs the light image. Using projection mode, the light image is transmitted through a sophisticated system of high quality optical lenses, delivering a sharper light image than can be attained in proximity mode (15).

The second limitation stems from the insolubilization mechanism operant in these resists. Photoinitiated cross-linking converts the polymer film

from a glassy matrix of individual polymer strands into a cross-linked three-dimensional network that is effectively a single molecule bound to the substrate surface. Cross-linked and unexposed regions of the film are structurally similar and interact comparably with the developer solvent. The same intermolecular forces that are the basis for dissolution of the unexposed film can lead to considerable permeation of the developer into the exposed and crosslinked portions. The resulting swelling distorts the resist images, in unfavorable cases so severely that adjacent resist features contact and adhere.

Both solvent-induced swelling and oxygen inhibition are characteristic of all cross-linking negative resists based on free-radical chemistry.

Dry-Film Resists Based on Radical Photopolymerization. Photoinitiated polymerization (PIP) is widely practiced in bulk systems, but special measures must be taken to apply the chemistry in lithographic applications. The attractive aspect of PIP is that each initiator species produced by photolysis launches a cascade of chemical events, effectively forming multiple chemical bonds for each photon absorbed. The gain that results constitutes a form of "chemical amplification" analogous to that observed in silver halide photography, and illustrates a path for achieving very high photosensitivities.

One limitation of PIP is that polymerizable monomeric species almost invariably are liquids or low melting solids. Films containing significant quantities of monomer are tacky even when combined with a polymeric binder, and so are susceptible to contamination-related defects. In 1968, E. I. du Pont de Nemours & Company introduced an innovative PIP resist system (marketed under the trade name Riston), supplied in the form of a multilayered "dry film" structure (2,11). The polymerizable layer is sandwiched between a polyolefin carrier sheet and a transparent polyester cover sheet (Fig. 7a). The resist is applied by removing the carrier sheet and laminating the polymerizable layer to the substrate using heat and pressure. The cover sheet is left in place, both to protect against contamination and to act as a diffusion barrier against atmospheric oxygen during exposure. Prior to developing the image the cover sheet is peeled away. Today there are a number of commercial resist products marketed in dry-film format.

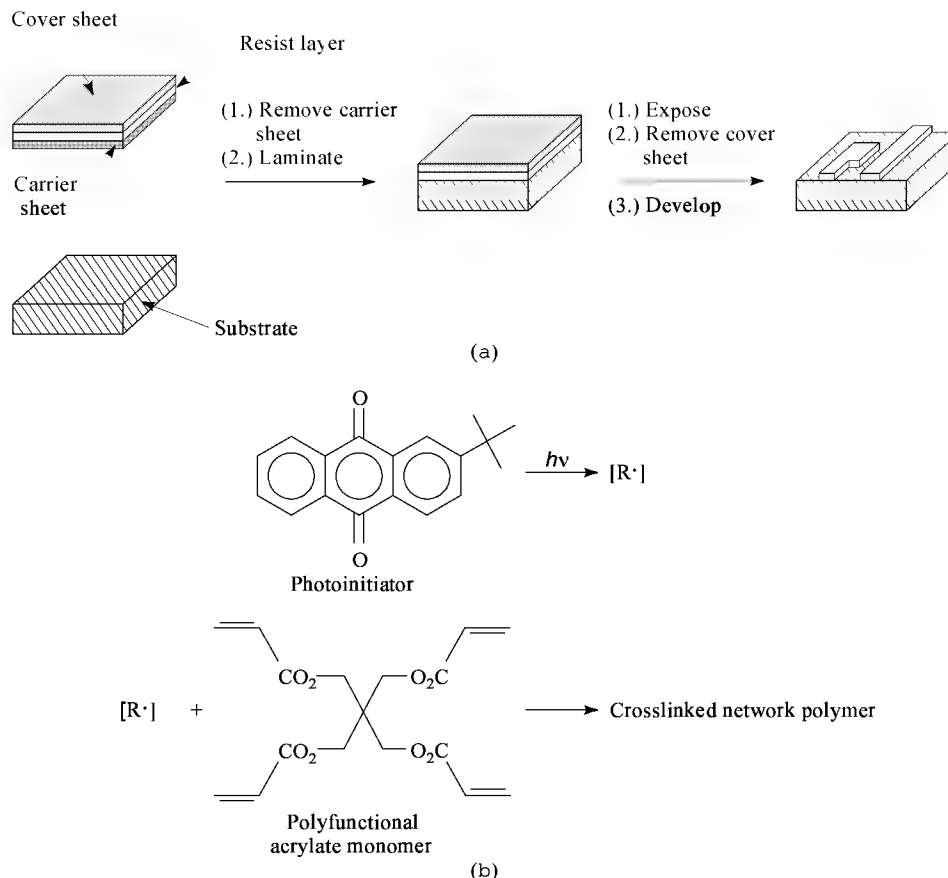


Fig. 7. Chemistry of a dry-film, negative-acting photoresist. (a) Polymerizable layer sandwiched between a polyolefin carrier sheet and a polyester cover sheet. (b) Chemical structures of typical components.

In resists of this class, the imaging layer contains a multifunctional monomer that can form an interconnected network upon polymerization, and a photosensitizer to generate a flux of initiating free radicals. Although not strictly required for imaging, the composition usually includes a polymeric binder (typically an acrylic copolymer) to modify the layer's physical properties. Figure 7b shows the chemical structures of typical components.

An important practical advantage of the dry film format is that the lamination process readily produces highly uniform films of precise thickness, avoiding solvent handling, post-apply baking, and the extreme cleanliness required when a resist film is coated from solution. The process simplification that results has led to the wide use of dry-film resists for the fabrication of printed wiring boards (PWBs) (11). The dry film format also provides a means for sealing and protecting holes or recesses in the substrate during image transfer. This property, termed *tenting*, is exploited in the fabrication of multilayered PWBs, where plated through-holes serve as electrical conductors to join levels of circuitry (Fig. 8).

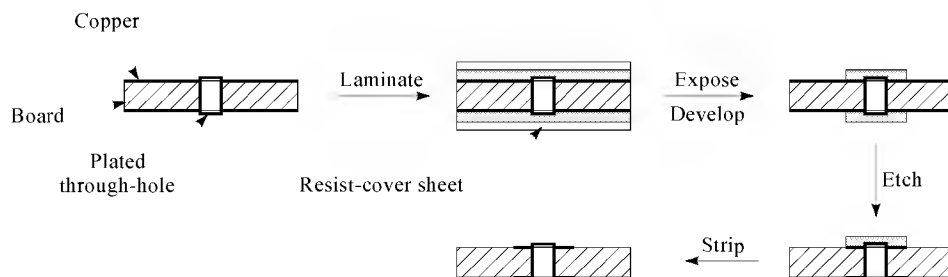


Fig. 8. Protection of a plated through-hole in a PWB during etching by using a tenting dry-film photoresist.

As PWB technology is refined to provide greater integration using finer conductor lines, there is renewed interest in liquid resists. The absence of a cover sheet and the ability to apply thinner films both contribute to improved resolution and to an intrinsically lower consumables cost (16,17).

Resist Materials for Imaging with Ionizing Radiation. Although most lithographic processing uses light in the visible to uv wavelength region (known as photolithography), a number of specialized, low volume applications employ high energy radiation and have unique resist

requirements as a result. These include very high resolution lithography techniques using beams of electrons or x-rays to induce chemical changes in the resist film. Electron beam or e-beam lithography is used extensively for the fabrication of high resolution photomask patterns, and for direct writing or circuit patterns for low volume, custom microelectronic applications (18). The applicability of proximity-mode x-ray lithography in large-scale device fabrication has been extensively researched (19). To date, however, photolithography has been extended to image resolutions well beyond those first anticipated a decade ago, deferring a conversion to x-ray tooling. At present the principal commercial application of x-ray lithography is the LIGA (a German acronym for lithography–electroplating–molding) process for forming micromechanical parts (6).

In e-beam lithography, a finely focused beam of high energy electrons is scanned in a pattern across a resist-coated substrate. Collisions of impinging (primary) electrons and secondary electrons with atoms of the resist film produced ionic and radical species that undergo further reaction, modifying properties of the surrounding matrix. In x-ray lithography, initial absorption of a photon produces an energetic photoelectron whose collisions with atoms of the resist lead to the same intermediates produced by e-beam radiation. For this reason there tends to be a correlation between the sensitivity of a resist material for electron beam and x-ray lithography. However, the chemistry differs substantially from that caused by nonionizing (ie, long wavelength ultraviolet to visible) radiation, so that there is no assurance that a resist designed for photolithography will be found suitable for these methods.

Many examples of positive- and negative-tone x-ray and e-beam resists have been described (1,20,21). A number of these are single-component systems consisting of a polymer with intrinsic radiation sensitivity. If radiation-induced cross-linking is the predominant reaction, then the resist is negative-acting. Conversely, if chain scission and/or depolymerization are the principal reactions, then the polymer functions as a positive-tone resist. One representative example is a positive-tone resist developed by Bell Laboratories in the early 1970s that has been widely applied to commercial photomask manufacture by electron-beam lithography (22). This resist is the polymer poly(butene-1-sulfone) (PBS), an alternating copolymer of butene and sulfur dioxide (Fig. 9). Electron-beam irradiation of PBS is believed to produce a cation radical center initially that destabilizes the chain. Scission at that site creates radical- and cation-terminated polymer strands which may undergo further fragmentation before deactivation.

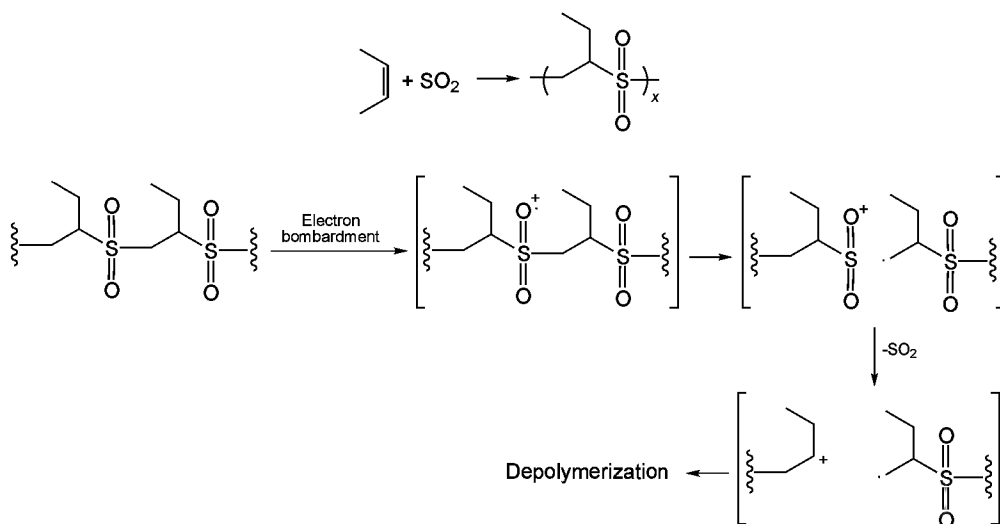


Fig. 9. Chemistry of PBS electron beam resist.

Modern Resists for Microlithography

Today the market for photoresists used in the manufacture of integrated circuits is estimated to be \$750 million U.S. annually, with diazonaphthoquinone–novolac resist products accounting for almost all of this. Newly introduced commercial resist systems designed for deep-ultraviolet applications are now being used to build leading edge semiconductor products, however, and forecasts suggest that use of such advanced materials will grow several-fold over the next five years.

Positive-Tone Photoresists based on Dissolution Inhibition by Diazonaphthoquinones. The intrinsic limitations of bis-azide–cyclized rubber resist systems led the semiconductor industry to shift to a class of imaging materials based on diazonaphthoquinone (DNQ) photosensitizers. Both the chemistry and the imaging mechanism of these resists (Fig. 10) differ in fundamental ways from those described thus far (23). The DNQ acts as a dissolution inhibitor for the matrix resin, a low molecular weight condensation product of formaldehyde and cresol isomers known as novolac (24). The phenolic structure renders the novolac polymer weakly acidic, and readily soluble in aqueous alkaline solutions. In admixture with an appropriate DNQ the polymer's dissolution rate is sharply decreased. Photolysis causes the DNQ to undergo a multistep reaction sequence, ultimately forming a base-soluble carboxylic acid which does not inhibit film dissolution. Immersion of a patternwise-exposed film of the resist in an aqueous solution of hydroxide ion leads to rapid dissolution of the exposed areas and only very slow dissolution of unexposed regions. In contrast with crosslinking resists, the film solubility is controlled by *chemical* and *polarity* differences rather than molecular size.

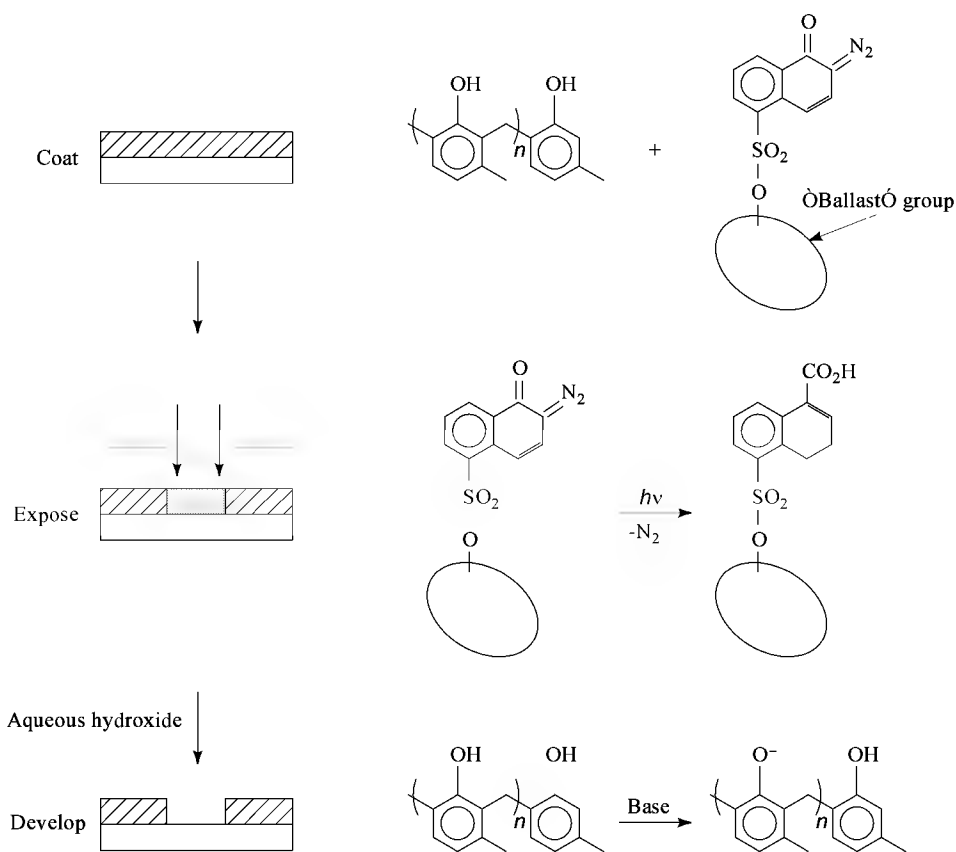


Fig. 10. Chemistry of diazonaphthoquinone-cresol novolac positive-acting resist. During patternwise exposure, the DNQ undergoes photolysis that destroys its inhibitory effect on film dissolution.

DNQ-novolac resists exhibit several important practical attributes. First, DNQ photochemistry is not inhibited by oxygen, so unlike free-radical based systems the resist can be exposed in noncontact modes with unimpaired imaging properties. Second, films of the resist dissolve in aqueous base by a surface-limited etching reaction, with no evidence of swelling. Presumably this is because only the ionic, deprotonated phenolate form of the polymer is sufficiently polar to be solubilized by the aqueous developing solvent. Finally, the developing solvent is nonflammable and water-based, an advantage from the viewpoints of workplace safety and materials handling and disposal.

DNQ-novolac resist chemistry has proved to have remarkable flexibility and extendibility. First introduced for printing applications, DNQ-novolac resists have been available since the early 1960s in formulations intended for electronics applications. At present, most semiconductor manufacturing processes employ this resist chemistry. Careful contemporary research and engineering support the continuing refinement of this family of materials.

DNQ Synthesis and Properties. DNQ photosensitizers are synthesized by base-catalyzed condensation of a diazonaphthoquinone sulfonyl chloride with a mono- or polyhydroxy species to produce a sulfonate ester (Fig. 11). Since the structures of the photolabile diazonaphthoquinone group and the photoinert ballast can be readily and independently changed using this route, many DNQ variants have been prepared. Modification of the diazonaphthoquinone group in large part has been restricted to positioning the sulfonate functionality on either the 4- or 5-location. The location of the sulfonate influences the chromophore's absorption properties, with the long-wavelength absorption ($\lambda_{\max}$ of ~ 380 nm) of the 4-sulfonate blue-shifted in comparison to the 5-sulfonate isomer ($\lambda_{\max}$ of ~ 400 nm). More significantly, the substitution pattern can influence the course of the photoinduced reaction. Like the 5-sulfonate isomers, a 4-sulfonate DNQ first undergoes a Wolff rearrangement to form a carboxylic acid upon photolysis. This initial product is labile, however, and the 4-sulfonate analogue can undergo further facile hydrolysis, forming a strongly acidic sulfonic acid as the final product (Fig. 12). This has been exploited in the design of an "image-reversing" resist, where sulfonic acid catalyzes an insolubilizing reaction if an exposed film is heated (25).

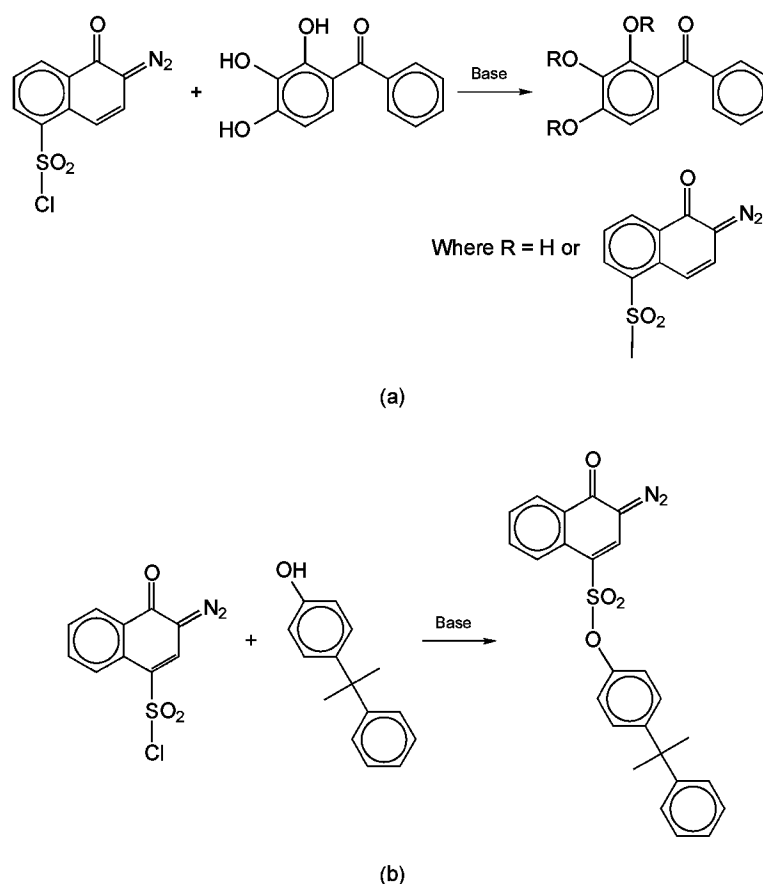


Fig. 11. Synthesis of DNQ photosensitizers found in commercial resists. (a) Condensation of 1,2-diazonaphthoquinone-5-sulfonyl chloride with 1,2,3-trihydroxybenzophenone. Often the reaction is not carried to completion so the product is a mixture of monodi- and trisubstituted products. (b) Condensation of 1,2-diazonaphthoquinone-4-sulfonyl chloride with *p*-cumylphenol.

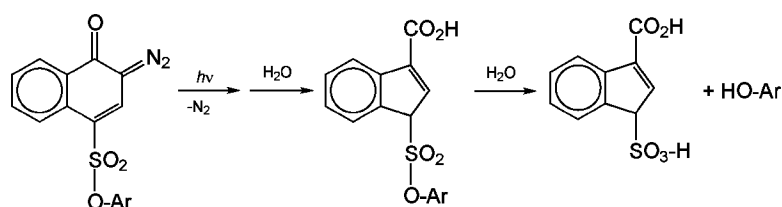


Fig. 12. Photo-induced chemistry of a 4-sulfonyl DNQ. The intermediate species reacts with adventitious water in the resist film to produce a sulfonic acid and a phenol.

Structural variation of the ballast or backbone group has been explored in far greater depth. By correlating structure and lithographic properties, researchers have identified numerous factors that can influence resist performance. These include the number of diazonaphthoquinone sulfonate groups per sensitizer molecule (26), the spatial proximity of those groups (27), the backbone's size and the degree of esterification of polyfunctional DNQs (28,29), backbone hydrophobicity (28), and DNQ's capacity for hydrogen-bonding with the phenolic matrix resin (30).

Novolac Synthesis and Properties. Novolac resins used in DNQ-based photoresists are the most complex, the best-studied, the most highly engineered, and the most widely used polymers in microlithography. Novolacs are condensation products of phenolic monomers (typically cresols or other alkylated phenols) and formaldehyde, formed under acid catalysis. Figure 13 shows the polymerization chemistry and polymer structure formed in the step growth polymerization (31) of novolac resins.

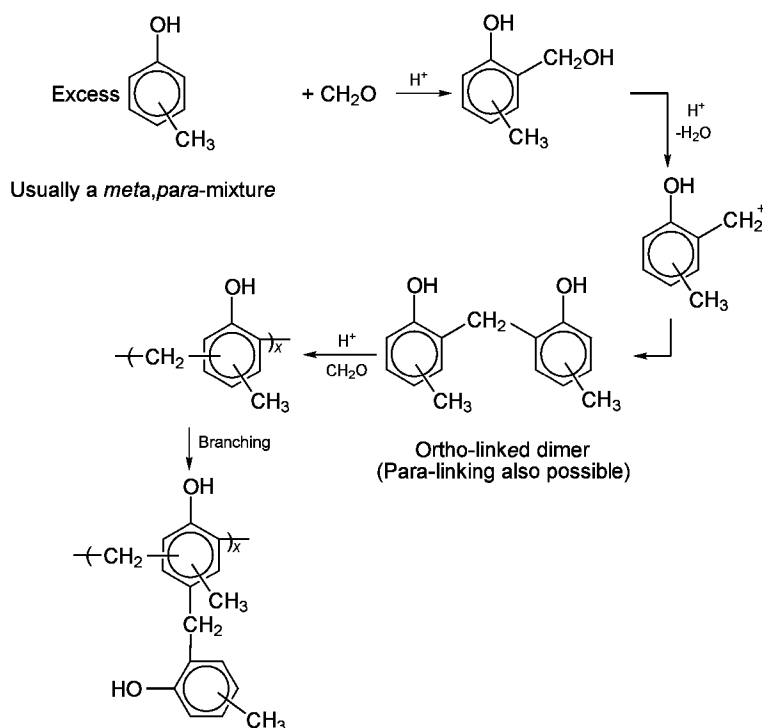


Fig. 13. Polymerization chemistry of phenol-formaldehyde condensation: synthesis of novolac resin. The phenol monomer(s) are used in stoichiometric excess to avoid gellation, although branching invariably occurs due to the multiple reactive sites on the aromatic ring.

The polymerization mechanism is thought to involve electrophilic attack on the aromatic ring at the positions ortho and para to the phenolic hydroxyl group (23). Initially, formaldehyde adds to the ring, forming a benzyl alcohol. In the second step, protonation of the alcohol and loss of a water molecule produces a benzyl cation that can undergo electrophilic substitution with another aromatic ring to form a dimeric product. Repeated reaction of the products with formaldehyde initiates a sequence of condensations to generate novolac polymer.

Molecular Weight Effects. In step-growth polymerization, high molecular weight product is formed only when the following two requirements are met: (1) strictly difunctional monomers are used in a 1 : 1 stoichiometric ratio, and (2) the extent of conversion of monomer to polymer approaches 100% (31). In such a case, classical step-growth kinetics would predict a most probable distribution of molecular weights with a polydispersity of 2.0. In the case of novolac synthesis, however, the phenolic monomer is not perfectly difunctional, but more likely should be considered trifunctional. One result of this "excess" functionality is that branching occurs early in the polymerization (Fig. 13). If the reaction is brought to a high degree of conversion, gellation and cross-linking result; the product is unusable (and difficult to remove from the reactor, as these polymerizations are often run in the bulk using no polymerization solvent). To avoid this, the polymerization is run with excess phenol and is not brought to high conversion.

The combined effects of branching, which leads to the formation of high molecular weight products, and the low overall extent of conversion, which favors formation of low molecular weight oligomers, yield a polymerization product of some complexity. As synthesized, novolacs have a very low average molecular weight, with a multimodal molecular weight distribution. Figure 14 shows a size-exclusion chromatogram of a typical novolac resin, illustrating the multimodal molecular weight distribution characteristic of this class of polymer.

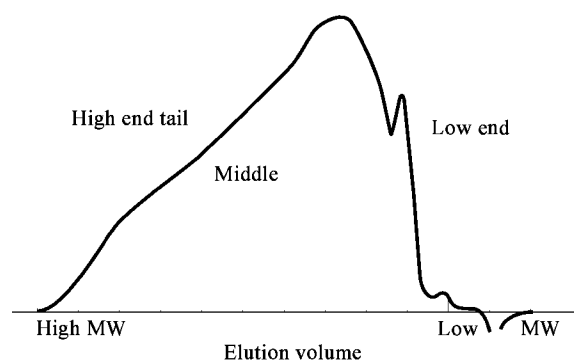


Fig. 14. Molecular weight characteristics of novolac resins. Shown is the size-exclusion chromatogram for a typical commercial novolac polymer. The unsymmetrical peak shape reflects the multimodal molecular weight distribution of the polymer.

Figure 15 graphically shows the well-known general relationship between polymer properties and molecular weight. Most polymers in practical use inhabit the plateau region of this curve, where properties change only slightly with molecular weight. Novolac resins, however, reside on the early portion of this curve, where a slight change in molecular weight strongly influences material properties. The broad molecular weight distribution, coupled with low average molecular weight, ensures that portions of the novolac mixture exhibit the full range of properties. In Figure 14, three main molecular weight regions can be distinguished: a region that contains small oligomers, a high molecular weight tail with $M_n > 4000$, and a midrange portion between these extremes. Isolation and examination of each of these individual components, a difficult task by traditional methods, has recently been achieved using supercritical fluid fractionation (32). This study showed that a typical novolac contains 25–50% high molecular weight polymer that in isolated form is insoluble in aqueous alkaline developer solutions, and 15–25% oligomeric liquid that dissolves essentially instantaneously in alkaline developer. Examination of materials properties of the near-monodisperse novolac fractions demonstrates the pronounced influence of molecular weight. For example, thermal analysis of the separated novolac fractions reveals a substantial change in glass-transition temperature (from below room temperature to above 160°C) depending on whether the low or the high end of the molecular weight distribution is chosen.

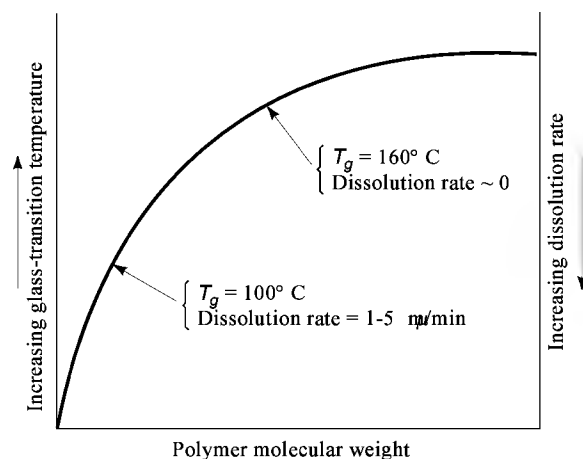


Fig. 15. Properties of novolac fractions are a strong function of molecular weight.

As one might anticipate, lithographic properties of DNQ-novolac resists are strongly influenced by polymer molecular weight and molecular weight distribution. For example, in one study, dissolution rates of both exposed (R_e) and unexposed (R_u) resist films were shown to decrease with increasing molecular weight, though the ratio of these rates R_e/R_u , a measure of contrast, was found to be independent of molecular weight (33). This same study also found that molecular weight distribution influences resist contrast, which improves with increasing polydispersity until a maximum contrast is attained. In recent work, a series of experimental DNQ-based resist formulations were prepared by systematically blending solvent-fractionated (34) or supercritical fluid fractionated (32) novolacs. By comparing lithographic imaging results for all the polymer mixtures, the influence of individual fractions on photospeed, contrast, and post-development residue and resistance to thermal flow were established. From this and related work has emerged a set of design principles that facilitate more precise optimization of resin molecular weight in resist formulations.

Structural Effects. The molecular structure of novolac resins can be readily manipulated by varying the mixture of phenolic monomers in the polymerization feed, and by selecting a different polymerization catalyst. Novolacs used in DNQ-based resists often are prepared from a mixture of cresol isomers, in particular a mixture of *para*- and *meta*-cresols, rather than a single component. The polymerization chemistry is sufficiently general and flexible such that a range of phenolic homologues and derivatives have been incorporated into novolacs with the goal of improving lithographic properties (23).

Numerous studies have probed how novolac microstructure influences resist lithographic properties. In one example, a series of resists were formulated from novolacs prepared with varying feed ratios of *para*-/*meta*-cresol. These researchers found that the dissolution rate decreased, and the resist contrast increased, as the *para*-/*meta*-cresol feed ratio increased (33). Condensation can only occur at the *ortho* position of *para*-cresol, but can occur at both the *ortho*- and *para*-positions of *meta*-cresol. It is believed that increased steric factors and chain rigidity that accompany increased *para*-cresol content modify the polymer solubility.

Deep-Ultraviolet Chemically Amplified Resists Based on Acid Catalysis. In any optical imaging system, the size of the smallest element that can be accurately resolved is related to the wavelength of exposing light: the smaller the wavelength, the finer the feature that can be resolved (35). In the microlithographic arena, improved resolution is achieved by incrementally shifting the exposure wavelength to smaller values as refinements in optics, tooling and process technology permit. The microelectronics industry currently is beginning a transition from exposure tools designed to use monochromatic light at $\lambda = 436$ and 365 nm (two strong emission lines in the spectrum of mercury arc lamps) to more advanced exposure tools that use light at 248 nm (the output wavelength of a krypton fluoride excimer laser), the deep-ultraviolet or duv region. This wavelength shift has profound implications from the viewpoint of resist design. DNQ-novolac resists, now the industry standard, are impractical for duv photolithography for two reasons.

The first can best be illustrated by consideration of resist optical properties. Figure 16 shows the optical absorption spectrum of a typical DNQ-novolac photoresist film before and after exposure with 365 nm light. The absorption band at long wavelengths (300–450 nm) is that of the DNQ sensitizer. As the DNQ is photolyzed this characteristic absorption decreases. The increased transparency facilitates efficient photolytic conversion of the sensitizer throughout the depth of the film. If such a film is exposed at 248 nm, however, the strong nonbleaching absorbance at the wavelength (due in part to the novolac matrix polymer, and in part, to the DNQ and its photoproduct) sharply attenuates the beam as it passes through the film. The result is highly nonuniform photolysis of photosensitizer in the resist film, with insufficient conversion near the substrate interface. Figure 17 illustrates the damaging effect on resist imaging properties.

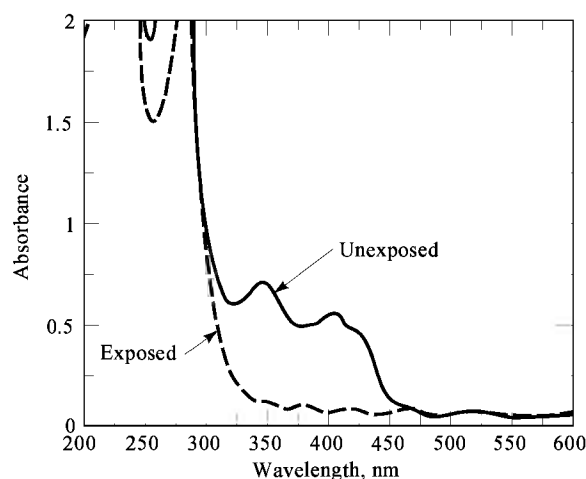


Fig. 16. Optical transmittance spectrum of a typical DNQ-novolac resist. (—), The transmittance of a 1 μm film after coating. (---), the transmittance of the film after exposure with light of 365 nm wavelength. Note the sharp increase in transparency at the exposure wavelength upon exposure, and the high opacity of the film at wavelengths below 300 nm.

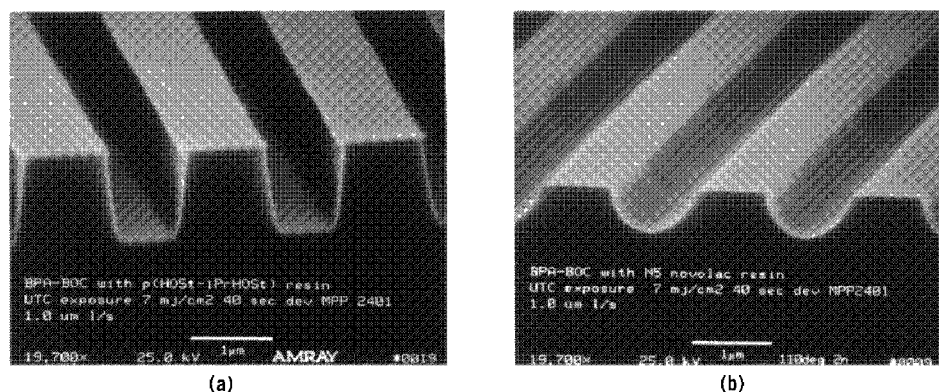


Fig. 17. Influence of optical absorbance on photoresist imaging properties. Shown are scanning electron micrographs of relief images obtained using two different photoresist formulations, processed under identical conditions. The formulations are functionally identical but differ in the transparency of the film at the exposing wavelength (36). (a) Image obtained when the film is weakly absorbing ($\sim 65\%$ transmittance). (b) Image obtained when the film absorbs strongly ($<20\%$ transmittance) at the exposing wavelength. Note the incomplete development to substrate and the sloping sidewalls.

Second, the overall brightness of light sources available for duv exposure is much less than that of mercury arc lamps used for 365 nm wavelength. Though KrF excimer lasers are generally regarded as powerful sources of uv light, the spectral output line is relatively broad. Limitations of available lens materials make correction for chromatic aberration difficult, so the source output beam must span only a very narrow wavelength range. Introduction of optical line narrowing elements into the beam leads to a large overall attenuation. Roughly speaking, a photoresist designed to be used with duv exposure tooling now at hand must be at least tenfold more efficient in its utilization of absorbed photons than a DNQ–novolac resist. Since the quantum yield for DNQ–novolac systems is on the order of ~ 0.3 , a $10\times$ improvement in efficiency is a considerable challenge.

One potential approach extends the idea of chemical amplification introduced in our preceding description of dry-film resists. In 1982, Ito and co-workers (37,38) recognized that if a photosensitizer producing an acidic product is photolyzed in a polymer matrix containing acid-labile groups, the acid will serve as a spatially localized catalyst for the formation or cleavage of chemical bonds.

An early practical example of this concept is shown in Figure 18. The acid-generating sensitizer is an onium salt, a class of compounds that efficiently generate a Brønsted acid upon photolysis. The polymer is poly(*t*-butoxyoxycarbonylstyrene) (PTBOCST). Photolysis forms a small quantity of acid in those areas exposed to radiation. During a later heating step the acid catalyzes thermolysis of pendent *t*-butoxyoxycarbonyl (TBOC) groups, converting the nonpolar PTBOCST into the polar poly(hydroxystyrene) (PHOST) and gaseous products, while regenerating the initial acid. The converted latent image can be developed in either positive tone by selecting a polar solvent, or in negative tone by selecting a nonpolar solvent as the developer. The catalytic chain length for this system, the average number of TBOC groups fragmented by each acid molecule, has been measured to be in the range 800–1100 (39). Therefore only extremely low exposure doses (in the range of $1\text{--}5\text{ mJ cm}^{-2}$) are needed to image this system, which is roughly two orders of magnitude more photosensitive than typical DNQ–novolac resists.

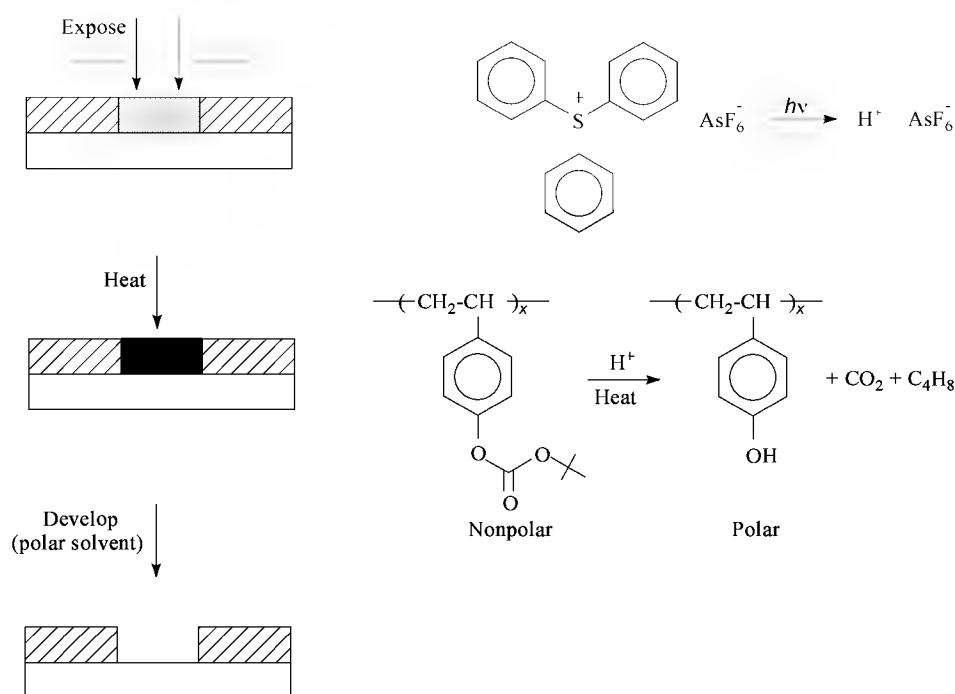


Fig. 18. Chemistry of PTBOCST chemically amplified resist. Patternwise exposure creates small quantities of acid. In a subsequent heating step, pendant TBOC groups are cleaved under acid catalysis. The exposed and unexposed areas can then be differentiated on the basis of solubility.

Since the original proof of concept, and a later demonstration of its practical use in semiconductor manufacturing (40), applications and extensions of this concept have proliferated. In the following sections these systems are described in greater detail with emphasis on the resist formulation at a components level (41).

Photoacid Generators. Practical applications of photoacid generators (PAGs) have been actively pursued since onium salts were first reported to serve as photopolymerization catalysts in the 1970s (42). Onium salts are particularly useful as photoinitiators since they generate both Brønsted acids and free radicals and consequently can simultaneously initiate both cationic and radical cures. Owing to the high quantum efficiency for acid production, these ionic PAGs are well-suited for chemically amplified photoresist applications. Onium salts are one of several classes of photoacid generators. A variety of other PAG compounds (both ionic and nonionic) have been developed as researchers have sought to optimize the following key functional properties:

- (1) Quantum yield of acid generation
- (2) Photoacid characteristics

Acid strength, pK_a
 Acid volatility
 Acid diffusion length

- (3) Wavelength response
- (4) Solubility
- (5) Thermal stability
- (6) Manufacturing costs
- (7) Toxicity

In the following sections the properties of photogenerators of strong Bronsted acids and their use in microlithography are summarized.

Ionic Photoacid Generators. Ease of synthesis, high thermal stability, and good quantum yield have made sulfonium and iodonium salts the most widely used onium salts. Figure 19 depicts some representative examples.

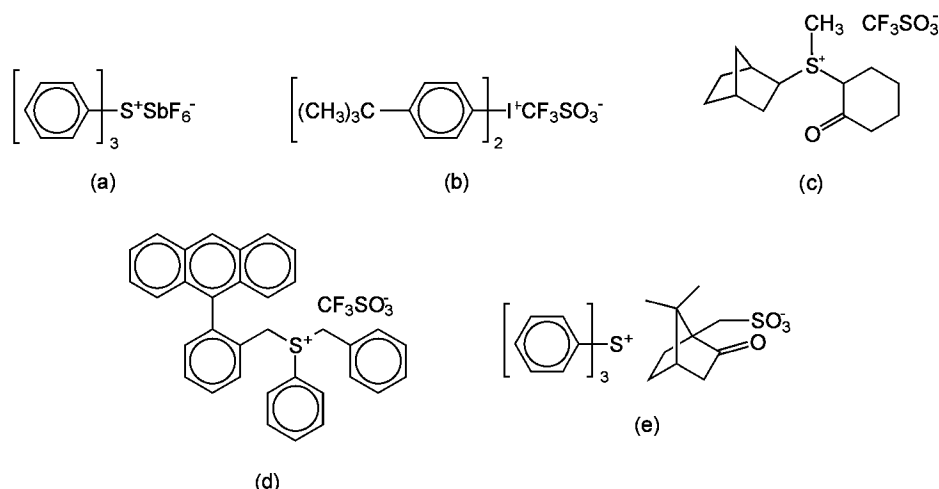


Fig. 19. Structures of selected sulfonium and iodonium salt photoacid generators designed for different lithographic applications. Modifications in the structure of the anion lead to changes in the properties of the photogenerated acid; changes in the structure of the cationic chromophore modify the light absorption properties of the PAG (43–47). (a) duv PAG generating inorganic acid; (b) duv PAG generating organic acid; (c) PAG designed for 193 nm lithography; (d) PAG designed for long wavelength activity via intramolecular electron transfer sensitization; and (e) PAG designed to generate bulky, weak acid.

By altering the counter ion in these salts (generally accomplished by a metathesis reaction) the properties of the photoacid can be readily modified, often without significant impact on the photochemical reactivity of the cation. In this way, the acid strength of the photogenerated acid can readily be varied from that of so-called "super acids" to the comparatively weak organic acids (48). Other modifications, such as the use of sterically bulky counterions (eg, substituting camphorsulfonate for methanesulfonate anion), have been studied in attempts to reduce acid volatilization and diffusion (which presumably improves lithographic resolution) (47).

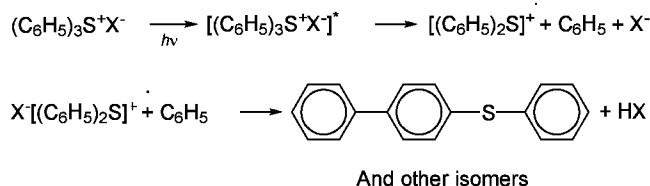
Diffusion of acid is believed to be an important factor determining CA resist performance. At the molecular level, diffusion is necessary to effect acid catalyzed reactions in the resist film. Diffusion also acts to smooth sidewall fine structure (resist standing waves) that results from interference effects within the resist film due to reflections from the film interfaces. (Many DNQ resists, which in principle require no thermal processing step after exposure, are heated after exposure to smooth out standing waves). However, if the mean diffusion length is too large the spatial profile of photogenerated acid is blurred and the final developed resist image appears distorted. The problem becomes more acute at finer resolution, and the characterization of acid diffusion and methods for its control in CA resists are active areas of research. The use of sterically bulky acids is one means for controlling photoacid diffusion, however, several other approaches are being pursued and are described in later sections.

The solubility properties of the PAG itself can play an important role in the overall resist performance as well (50). Solubility differences between the neutral onium salt and the acidic photoproducts can be quite high and will affect the resist contrast. In fact onium salts can serve as dissolution inhibitors in novolac polymers, analogous to diazonaphthoquinones, even in the absence of any acid-sensitive chemical function (51).

Simple aryl onium salts (see Fig. 19) above absorb in the wavelength range of 200–330 nm, making them well-suited for applications requiring duv exposure. Spectral response can be shifted to longer wavelengths either by varying the substituents (see Fig. 19) or by sensitization. Aromatic hydrocarbons with low oxidation potentials have been used to shift the activity of the PAG to wavelengths >500 nm via electron-transfer sensitization (52). The absorption properties can also be shifted to shorter wavelengths by employing aliphatic substituents in the PAG (see Fig. 19). These latter materials have been used in resists designed for exposure with 193 nm light, where aryl-substituted onium salts are claimed to increase absorption of the formulated resists to unacceptably high values (45).

The chemical pathways leading to acid generation for both direct irradiation and photosensitization (both electron transfer and triplet mechanisms) are complex and at present not fully characterized. Radicals, cations, and radical cations all have been proposed as reactive intermediates, with the latter two species believed to be sources of the photogenerated acid (Fig. 20) (53). In the case of electron-transfer photosensitization, aromatic radical cations (generated from the photosensitizer) are believed to be a proton source as well (54).

Direct irradiation



Photosensitization

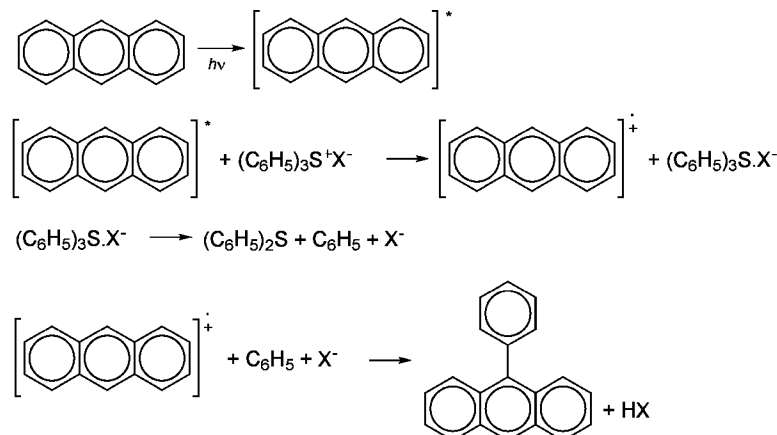


Fig. 20. Proposed photochemical mechanisms for the generation of acid from sulfonium salt photolysis. Shown are examples illustrating photon absorption by the onium salt (direct irradiation) as well as electron transfer sensitization, initiated by irradiation of an aromatic hydrocarbon.

Nonionic Photoacid Generators. Examples of such materials are listed in Figure 21, and include a variety of structural types which may undergo several different photochemical rearrangements. As one example, irradiation of the 2,6-dinitrobenzyl ester shown in Figure 21 results in an intramolecular rearrangement (the well-known *ortho*-nitrobenzyl rearrangement (55)) ultimately producing toluenesulfonic acid (56). In this case, no radical flux is produced upon photolysis, minimizing the possibility of cross-linking side reactions that can lead in some cases to resist scumming or even negative-tone behavior in positive resist systems (57). In another type of photochemical reaction, electron transfer from a photoexcited phenolic moiety of the polymer matrix to a ground (electronic) state PAG molecule (analogous to that depicted in Fig. 20) leads to the formation of methanesulfonic acid (Fig. 21) (58). Finally, 4-sulfonyl DNQ sulfonic acid generators, introduced earlier in the description of negative-tone novolac-based resists and shown in Figure 12, have been used as PAGs in CA resists, most effectively at exposure wavelengths longer than 300 nm.

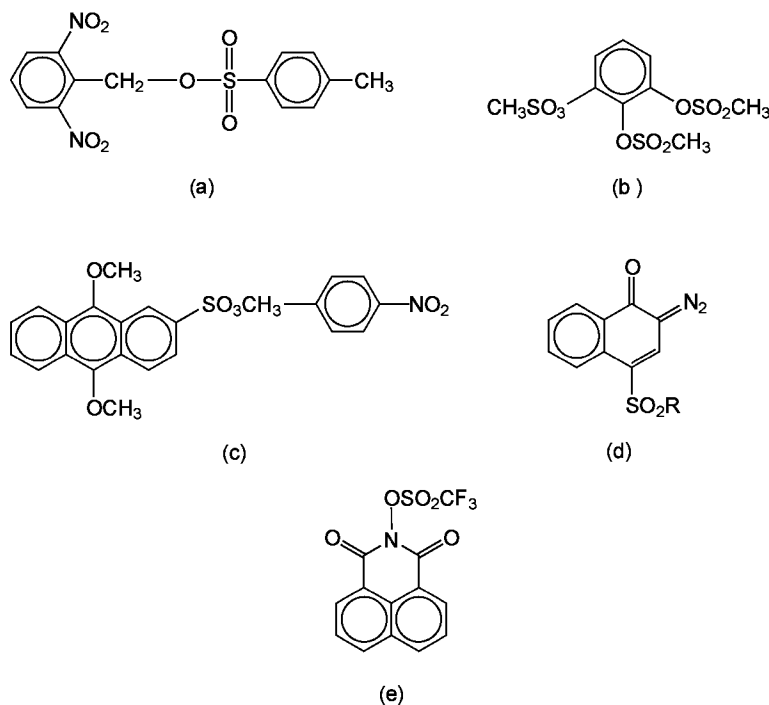
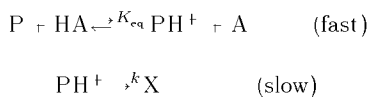


Fig. 21. Representative nonionic photoacid generators. A variety of photochemical mechanisms for acid production are represented. In each case a sulfonic acid derivative is produced (25,56,58–60). (a) PAG that generates acid via *O*-nitrobenzyl rearrangement; (b) PAG that generates acid via electron transfer with phenolic matrix; (c) PAG that is active at long wavelengths via electron-transfer sensitization; (d) PAG that generates both carboxylic acid and sulfonic acid; and (e) duv PAG generating trifluoromethane sulfonic acid.

To achieve the best overall resist performance, the optimum PAG for a given resist system, whether ionic or nonionic, must balance the functional properties listed earlier in this section. The development of new photoacid generators, and the characterization of their functional properties, are considered key to the design of resists with increased levels of performance.

Acid-Catalyzed Chemistry. Acid-catalyzed reactions form the basis for essentially all chemically amplified resist systems for microlithography applications (61). These reactions can be generally classified as either cross-linking (photopolymerization) or deprotection reactions. The latter are used to unmask acidic functionality such as phenolic or pendent carboxylic acid groups, and thus lend themselves to positive tone resist applications. Acid-catalyzed polymer cross-linking and photopolymerization reactions, on the other hand, find application in negative tone resist systems. Representative examples of each type of chemistry are listed below.

Positive-Tone Photoresists. The ester, carbonate, and ketal acidolysis reactions which form the basis of most positive tone CA resists are thought to proceed under *specific acid catalysis* (62). In this mechanism, illustrated in Figure 22 for the hydrolysis of *tert*-butyl acetate (type A_{Al}1) (63), the first step involves a rapid equilibrium where the proton is transferred between the photogenerated acid and the acid-labile protecting group:



The rate of the reaction in such case is $R = k[PH^+]$, where P is the reactant (ie, a repeat unit bearing the acid-labile protecting group).

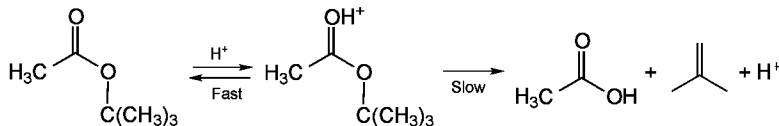


Fig. 22. Mechanism for the hydrolysis of *tert*-butyl acetate under strongly acidic conditions.

In resist systems the situation is more complex than that just outlined. The photogenerated acids can be so strongly acidic that they dissociate completely upon formation. In such case, the proton resides on other basic components of the resist film, which can include polymer functionality (ie, ester or ether comonomers), adventitious impurities such as water or residual casting solvents, as well as additives such as coating aids. The reactant PH^+ will be in equilibrium with all other protonated species in the resist film. The concentration of each protonated species is then determined by the basicity of its unprotonated conjugate. As deprotection proceeds, the concentrations of the basis resist components changes, shifting the equilibria and influencing the overall kinetics. Substitution of an acid of different strength, or introduction of a basic impurity or additive, perturbs the concentration of PH^+ , and thereby, the imaging behavior of the resist.

Figure 23 lists representative acid-labile protecting groups that have been incorporated in positive-tone CA resist systems. These groups can be pendent to the matrix polymer chain, can be attached to a monomeric or polymeric additive that acts as a dissolution inhibitor (64–66), or can even be appended to the PAG structure (67). The kinetics of acid-catalyzed deprotection vary significantly with structure. In particular, the activation energy, E_a , which characterizes how the rate of deprotection changes with temperature, has a large impact on CA-resist processing characteristics. For example, removal of the *tert*-butyl group from *tert*-butyl methacrylate has a relatively high E_a (measured to be 35 kcal/mole in experiments in resist films (68)), and resists based on this chemistry require postexposure processing at temperatures 120°C or higher. Since the extent of deprotection (and the related changes in the dimensions of the lithographic image) show a significant dependence on temperature, very precise temperature control is required. In contrast, resists based on low E_a chemistry (eg, the acetal system shown in Fig. 23) undergo deprotection at or near room temperature and exhibit little variation in final image linewidth with processing temperature (69).

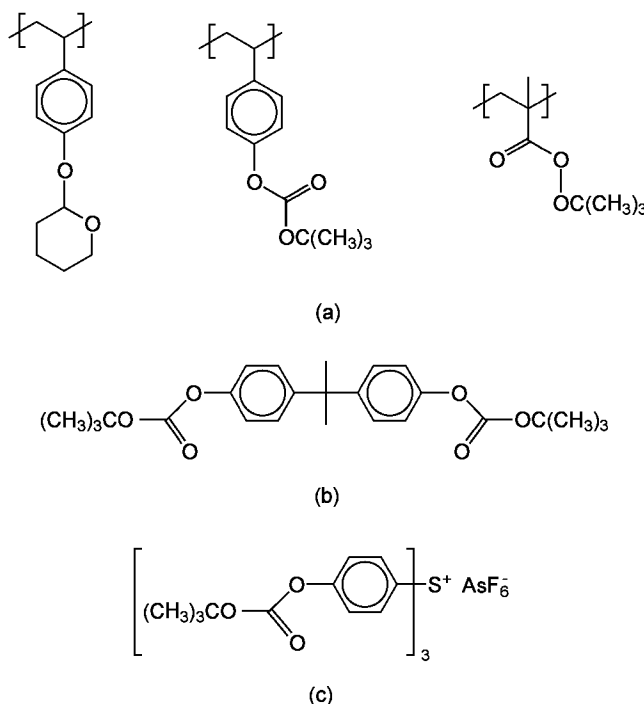


Fig. 23. Representative protecting groups for phenolic and carboxylic acid-based systems. (a) The polymer-based protecting groups are listed in order of increasing activation energy for acid-catalyzed deprotection. (b) Acid-labile monomeric dissolution inhibitors, a bifunctional system based on protected bisphenol A. (c) Another system that combines the function of dissolution inhibitor and PAG in a single unit.

Another potential advantage of these low activation energy systems is a reduced dependence on the precise timing of process steps (70–72) since there is in effect no delay between the exposure and post-bake steps. A potential disadvantage of these systems is that their intrinsically higher reactivity can reduce the stability of the formulated resist solutions during storage (ie, its shelf life). As with PAGs, in choosing a specific protecting group, a number of functional properties must be balanced to optimize overall resist performance.

Negative-Tone Photoresists. Many negative-tone CA resist employ acid-catalyzed cross-linking to effect insolubilization of the exposed pattern. These systems function in a manner analogous to radical-based cross-linking resists, in some cases exhibiting the same limitations. One of the first CA resist systems described in the chemical literature is based on the acid-catalyzed cross-linking of epoxy groups (Fig. 24a) (43). Although such resists display excellent photospeed and are immune to inhibition by molecular oxygen, they suffer from the same solvent-induced image distortion during development that is seen with radical crosslinking resists. Careful attention to formulation and process conditions is required to achieve high resolution imaging. These difficulties, coupled with the practical problems associated with the use of organic solvents in a manufacturing setting, have limited the acceptance of such resists.

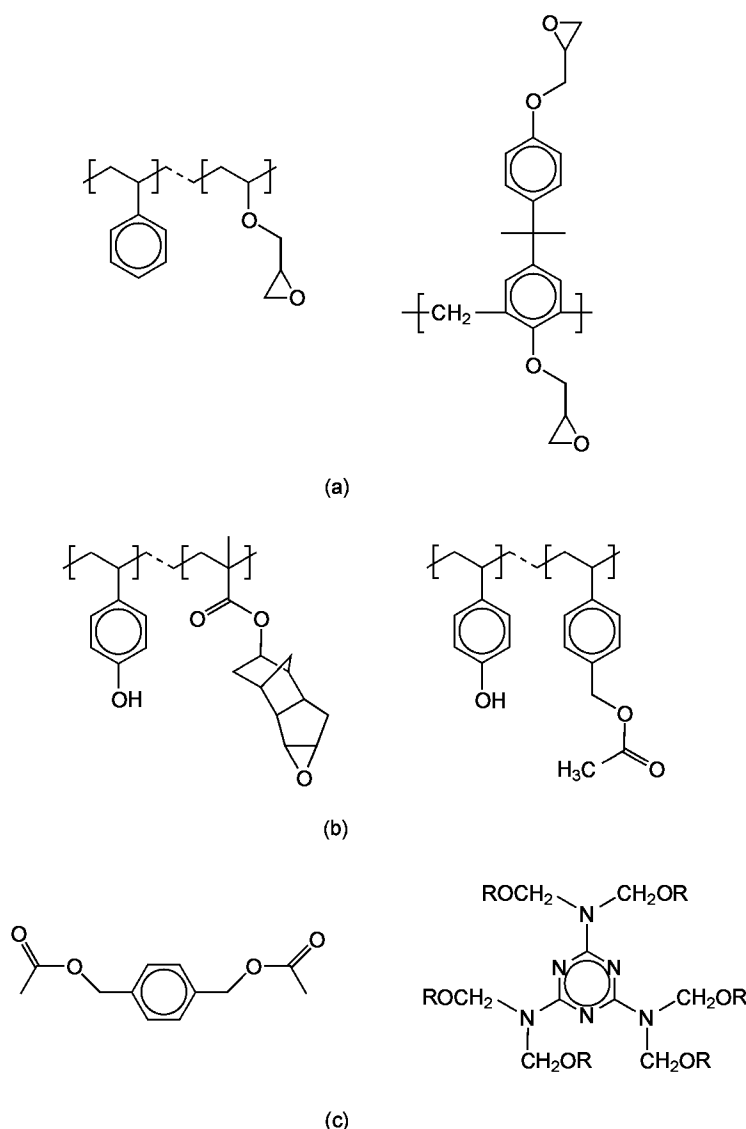


Fig. 24. Representative cross-linking systems employed in negative tone CA resists. (a) Epoxy polymers requiring organic solvent development. (b) PHOST-based cross-linking systems requiring aqueous development. (c) Monomeric cross-linking agents used in PHOST matrix polymers.

Several CA negative-tone photoresists based on polyhydroxystyrene (PHOST) polymers incorporate the cross-linking chemistries shown in Figure 24b and 24c. The phenolic structure renders these systems developable using aqueous alkaline solutions. Dissolution properties of this type of cross-linking resist closely resembles the surface limited etching found in DNQ–novolac positive resists, so image distortion due to solvent swelling is not observed. Systems have been described where the active cross-linking group is covalently bound to the polymer (73), or is incorporated as a nonpolymeric additive (74). Cross-linking can occur at the phenolic hydroxyl site or on the ring via electrophilic aromatic substitution reaction, as shown in Figure 25 (75).

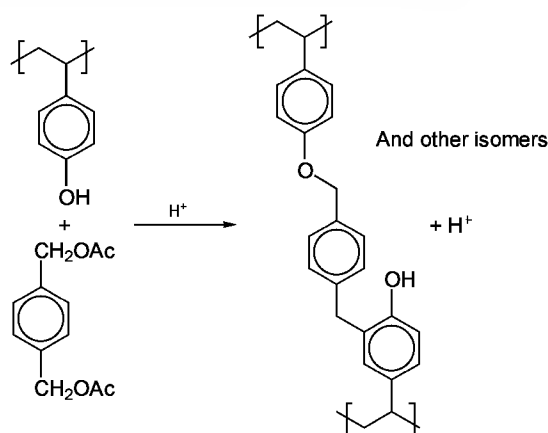


Fig. 25. Cross-linking reactions in a three-component resist system. Both O-alkylation and C-alkylation are shown.

Although cross-linking systems have been most widely studied, negative-tone patterning can be achieved by exploiting the polarity change that accompanies deprotection. For example, the TBOCST resist (see Fig. 18) produces negative-tone relief images when developed in a nonpolar organic solvent.

Airborne Basic Chemical Contamination. A critical, and at-first puzzling problem, was encountered during early manufacturing trials of CA resists. Sporadically, severely distorted resist profiles would be formed in positive-tone CA resists, displaying what seemed to be a cap on the upper surface of the resist image (Fig. 26). In severe cases this cap or T-top would appear as a kin or crust over the entire wafer surface that prevented development of the pattern. The magnitude of the effect varied dramatically between laboratories and appeared to grow more severe as the time interval between exposure and post-exposure bake was increased.

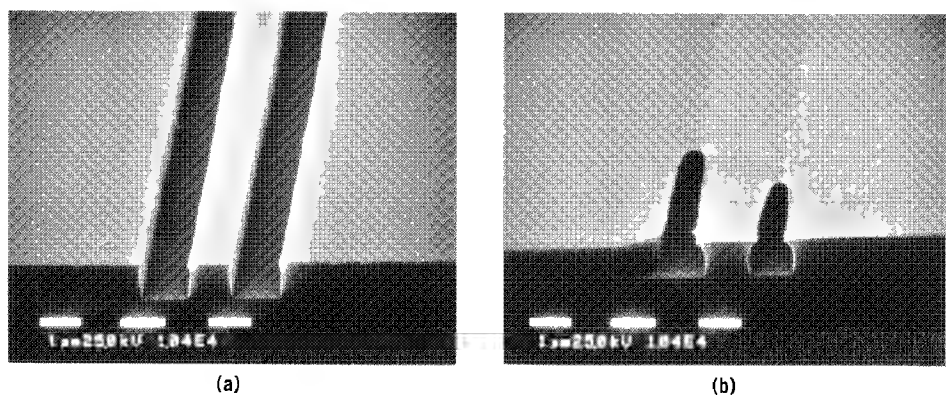


Fig. 26. SEM micrographs illustrating the effects of airborne basic chemical contamination. (a) This image was formed when a positive-tone CA resist was processed without any delay after coating. (b) This image was formed when an identical film was stored after coating for 15 minutes in an atmosphere containing 10 parts per billion on *N*-methylpyrrolidone, and then processed identically to the first film.

Ultimately, it was demonstrated that contamination of the resist surface by airborne basic impurities caused neutralization of the photogenerated acid near the resist–air interface, suppressing the acid-catalyzed chemistry, and thereby producing the observed effects. *N*-Methyl-pyrrolidone (NMP), a common solvent used in semiconductor manufacturing as a stripping and coating solvent, was the first basic contaminant to be identified in laboratory air (76,77). Even though an amide like NMP is generally considered to be a weak base (the pK_a of protonated NMP is ~ 1.0 (78)) when compared to the basicity of the polymer matrix (the pK_a of most protonated esters, ethers and phenols is in the range of -6 to -7), NMP is sufficiently basic to compete effectively for the proton catalyst, slowing the rate of the deprotection reaction (and the overall extent of conversion) in the contaminated region of the resist film. Other sources of airborne cleanroom contaminants now have been identified (76). These include paints, adhesives and sealants, materials that, like NMP, cannot be readily eliminated from the semiconductor manufacturing process.

The impact on negative-CA resists of airborne base contamination differs qualitatively from their positive tone counterparts. Suppression of acid-catalyzed chemistry at the surface of a negative resist results in some film erosion at the top of the exposed fields and in some cases an apparent loss of photosensitivity, but in general the relief images formed exhibit the expected cross-sectional profile. This is in sharp contrast with the typical behavior seen with positive-tone CA resists, where suppression of acid-catalyzed chemistry at the surface causes an insoluble surface skin.

Three approaches have been identified that reduce susceptibility of CA resists to airborne contamination. In the first, process engineering changes such as the addition of special activated carbon filters to the environmental chambers surrounding the exposure tools (76,79), overcoating the resist with a soluble protective film to isolate the resist from the environment (77,80,81), or modifications of the process flow to minimize the time interval between exposure and post-exposure bake have been shown to improve CA resist processibility.

A second approach modifies the CA resist chemistry. For example, researchers have introduced basic additives into the resist formulation to minimize the impact of surface contamination of the resist film (82,83). A resist that already contains added base (and consequently requires a larger imaging dose) should be less affected by the absorption of small amounts of basic contaminants. Systems of this type have been claimed to have improved resolution as well. The rationalization here is that the acid that diffuses into the unexposed regions of the resist film is neutralized and does not contribute to image degradation (84,85).

The third approach employs modifications of the polymer's physical properties and/or resist processing to minimize contaminant absorption, and is described in the section, "Polymer Properties and Lithographic Performance".

Polymers for Chemically Amplified Resists.

Derivatives of Polyhydroxystyrene. Earlier it was noted that novolac-based resists are too strongly absorbing at 248 nm to be used for duv exposure. Many derivatives of poly(4-hydroxystyrene) (PHOST) are sufficiently transparent at 248 nm to provide a phenolic platform for resist design. Figure 27 compares the optical absorption characteristics of PHOST and novolac polymers. Resists used in duv lithography are typically copolymers of hydroxystyrene and an acid-labile comonomer, often a protected hydroxystyrene, as depicted in Figure 28. These polymers are almost always produced by free-radical solution polymerization of a derivative of 4-hydroxystyrene (HOST), and thus, are usually medium molecular weight polymers with a relatively narrow polydispersity (between 1.5 and 2.5) (86). As detailed earlier, the unique properties of novolacs stem in part from the unusual condensation polymerization mechanism which produces substantial quantities of oligomeric products and high molecular weight branched polymer. Analogous fractions are not produced in the polymerization of protected hydroxystyrenes. The functional properties of phenolic resins based on HOST bear little resemblance to those of novolacs, even though the polymers are structurally similar (in fact, isomeric).

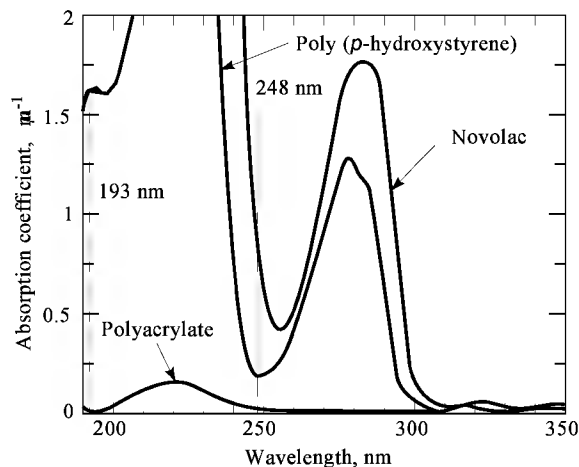
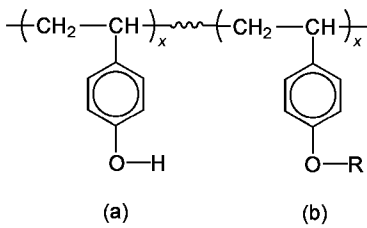


Fig. 27. Optical absorption spectra of thin, 1 μm -films of novolac, polyhydroxystyrene and polyacrylate polymers. The novolac resin is transparent only above 300 nm. While polyhydroxystyrene also absorbs strongly below 300 nm, it exhibits a region of adequate transparency centered near 248 nm. The acrylate polymer is quite transparent over the uv range.



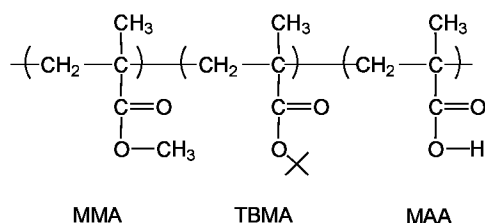


Fig. 31. An acrylic terpolymer designed for chemically amplified resist applications. The properties each monomer contributes to the final polymeric structure are for MMA, PAG solubility, low shrinkage, adhesion and mechanical, strength for TBMA acid-catalyzed deprotection; and for MAA, aqueous base solubility and increased T_g .

Incorporation of cyclic aliphatic (alicyclic) side groups markedly improves the plasma etch resistance of acrylic polymers, without reducing optical transparency at 193 nm (91). Figure 32 presents structures of some acrylic polymers currently under study for use in 193-nm CA resists (92–94). Recently, polymers with main-chain alicyclic structures have been described that offer similar properties (95,96).

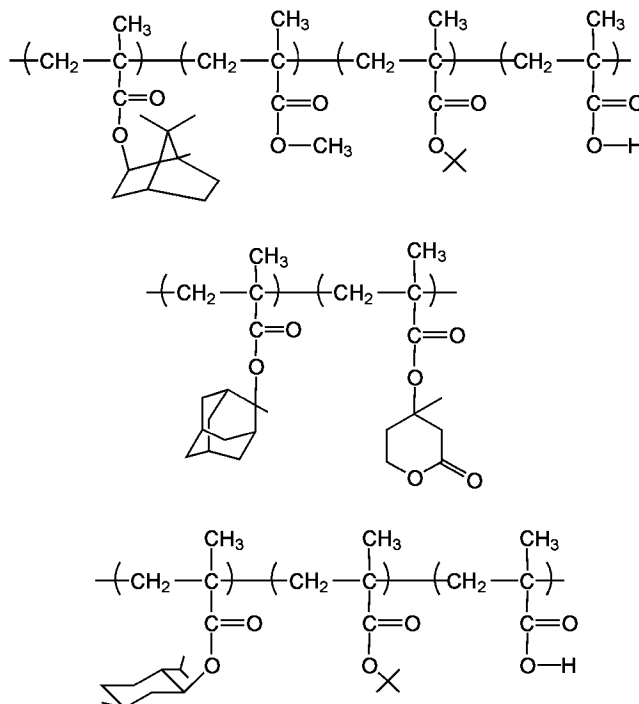


Fig. 32. Acrylic polymers found in CA resists designed for lithographic applications using 193 nm light.

Polymer Properties and Lithographic Performance in Chemically Amplified Resins.

Optical Absorption. Today, CA resists are being used in advanced duv photolithography applications where the exposure wavelength is between 190 and 250 nm. Most organic compounds absorb strongly in this spectral range, however, and the design of polymers for CA resists that maintain acceptable optical transparency while satisfying all other functional requirements is a challenge. The optimum optical absorbance for a nonbleaching photoresist is about 0.4 absorbance units per micron, representing a balance between conflicting requirements for high transparency to provide uniform illumination through the depth of the film, and high absorption to suppress thin-film optical reflectivity effects that reduce process latitude (97,98). To maximize light absorption by the photosensitizer, a rule of thumb is that the absorbance of the polymeric component should not exceed 0.2 per micron at the exposure wavelength.

Dry Etch Resistance. Gas-phase plasma etching image transfer processes (99) are ubiquitous in modern semiconductor fabrication. A polymer's stability in a plasma etch environment is a key consideration in resist design. Plasma etching processes evolved using DNQ–novolac resists so new lithographic materials must exhibit comparable etch resistance. Depending on structure and elemental composition, polymer plasma etch resistance can vary by orders of magnitude (100). A relationship has been established linking a resist's chemical composition with its resistance to a reactive ion etch (RIE) environment. Devised by Ohnishi and co-workers (101), a compositional parameter with good correlation to RIE rate is

$$\text{Ohnishi parameter} = N / (N_c - N_o)$$

where N , N_c , and N_o are the total number of atoms in the polymer repeat unit, number of carbon atoms, and number of oxygen atoms, respectively. This model serves as a fair predictor of etch rates for polymers under conditions where physical ion bombardment is a significant component, as is the case with RIE. The relation fails for low ion energy plasma conditions (eg, in a downstream glow discharge) where etching mechanisms are mainly chemical in nature. The Ohnishi parameter predicts that polymers having a high effective carbon content (low Ohnishi parameter) will exhibit low etch rates. Incorporation of oxygen and hydrogen into the repeat unit structure increases the etch rate, while an increased carbon content reduces the rate. Aromatic rings are, thus, quite etch resistant while aliphatic structures containing oxygen are etched rapidly. These concepts have been recently refined by introducing an additional parameter based not on composition but on polymer structure, termed the ring parameter (102). The added parameter extends the predictive capability of the relation to materials in plasma processes with low ion energies. This model recognizes that polymer etch rate is closely coupled to the fraction of the resist mass existing as or contained in ring structures, and can accurately predict the etch rate of many new alicyclic materials developed for 193-nm resists.

Thermal Properties. The glass-transition temperature of the polymer can influence CA resist properties in several ways. As with all resists, the final relief image must not exhibit significant thermal flow at the elevated temperatures reached during image transfer. With cross-linking systems, thermal flow is generally not a concern, but with positive-tone systems it can be problematic. Although the T_g of the base resist polymer may be high (eg, the T_g of PHOST homopolymer is 160–180°C), internal plasticization by the acid-labile group and external plasticization by resist additives and residual solvent will lower T_g of the final resist image (103).

Unlike DNQ–novolac resists, CA-resist imaging characteristics are determined to a significant extent by thermally activated bimolecular chemistry taking place during postexposure processing. Since the polymer serves here as the reaction medium, its properties and state influence the course and

evolution of chemical change. One particular concern is the impact of diffusion of the photogenerated acid on process latitude and on minimum achievable resolution. Several studies have probed how properties of the polymeric matrix influence diffusion in CA resists.

The impact of CA-resist thermal processing on acid mobility has been investigated, using a novel experimental method for generating sharply defined concentration gradients (104). In this work the mobility of photogenerated acid was seen to increase sharply in films as the temperature of the post-exposure bake was increased. The authors concluded that optimum lithographic performance is obtained when the post-exposure processing temperature is as low as possible, and in particular is below the T_g of the film. In contrast, acid mobility was shown to increase as the post-apply temperature was lowered. This effect was attributed to increased residual solvent content which facilitates acid migration.

In other work, the impact of thermal processing on linewidth variation was examined and interpreted in terms of how the resist's varying viscoelastic properties influence acid diffusion (105). The authors observed two distinct behaviors, above and below the resist film's glass transition. For example, a plot of the rate of deprotection as a function of post-exposure processing temperature show a change in slope very close to the T_g of the resist. Process latitude was improved and linewidth variation was minimized when the temperature of post-exposure processing was below the film's T_g .

Resistance to Airborne Chemical Contamination. Several methods for attenuating a CA resist's sensitivity to the presence of trace volatile bases were described earlier. Resist process stability in the presence of airborne bases also can be improved by consideration of how this stability is influenced by properties of the resist polymer. Using radiotracer methods to tag the basic contaminant, the kinetics and extent of contaminant uptake can be quantified (106). In one study, a wide range of polymer structures were cast as thin films on silicon wafers, and then stored in air containing a known, very low concentration of ^{14}C -labeled NMP (107). The rate of NMP uptake was found to vary by almost a factor of fifty depending on the polymer structure. The contaminant uptake of each polymer could be related to its physical properties, specifically its solubility parameter and T_g . The correlation between NMP uptake and polymer T_g (Fig. 33) was attributed to a reduction in free volume (a consequence of the nonequilibrium nature of the spin-coating process) due to polymer film annealing during post-apply baking, thereby reducing diffusant mobility.

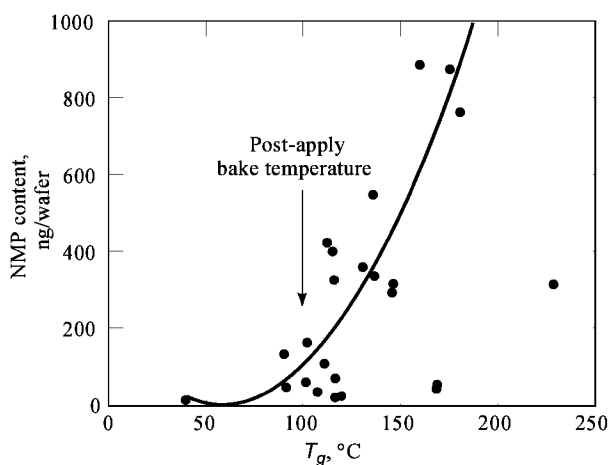


Fig. 33. Correlation of NMP uptake with polymer T_g for a series of 24 different polymeric films. All films were heated after coating at 100°C for 300 seconds (107).

One practical conclusion that can be drawn from this result is that susceptibility to airborne chemical contaminants can be minimized if a CA-resist film is baked after coating at a temperature at or above its T_g . However, at that time few existing resist systems had sufficient thermal stability to allow such treatment (108). By deliberately selecting polymer, protecting groups and PAGs with thermal stability in mind, however, CA resists can be designed that can be heated at or above T_g (109,110). Such resists display sharply improved environmental stability compared to analogous systems processed below T_g .

Extending the Chemically Amplified Resist Concept

Microlithography as used in modern semiconductor manufacturing has advanced largely through evolutionary refinement of exposure tool optics and mechanics and resist materials. The switch to duv-CA resists and excimer-laser-based exposure tools now underway in many respects represents a departure from this evolutionary process, and constitutes a considerable technical and financial challenge to the electronics industry. If continued improvements in lithographic performance are to be achieved in the future, it is clear that still more advanced technologies will be introduced (111). Figure 34 (112) lists several lithographic technologies being studied as extensions and replacements for duv technology when the scale of microelectronic devices shrinks to the 0.1 μm range. In parallel with the development of new exposure technologies utilizing high energy radiation and particle-beam sources, modifications and refinements of duv photolithography (248 and 193 nm indicated by a $\mp$ sign in Fig. 34) are being pursued. The simultaneous development of these technologies is extremely costly, so much of the research today is done in industry-government consortia worldwide. There is considerable uncertainty (and in some cases controversy (113,114)) over how future technology investments should be made since it is clear that the semiconductor industry will not be able to support simultaneously, the infrastructures required by all the technologies listed here. Although Figure 34 explicitly specifies only exposure technologies, the availability, properties and functions of suitable resist materials are key considerations when assessing the viability of these options. In the following section these new imaging technologies, and their potential impact on resist technology are described.

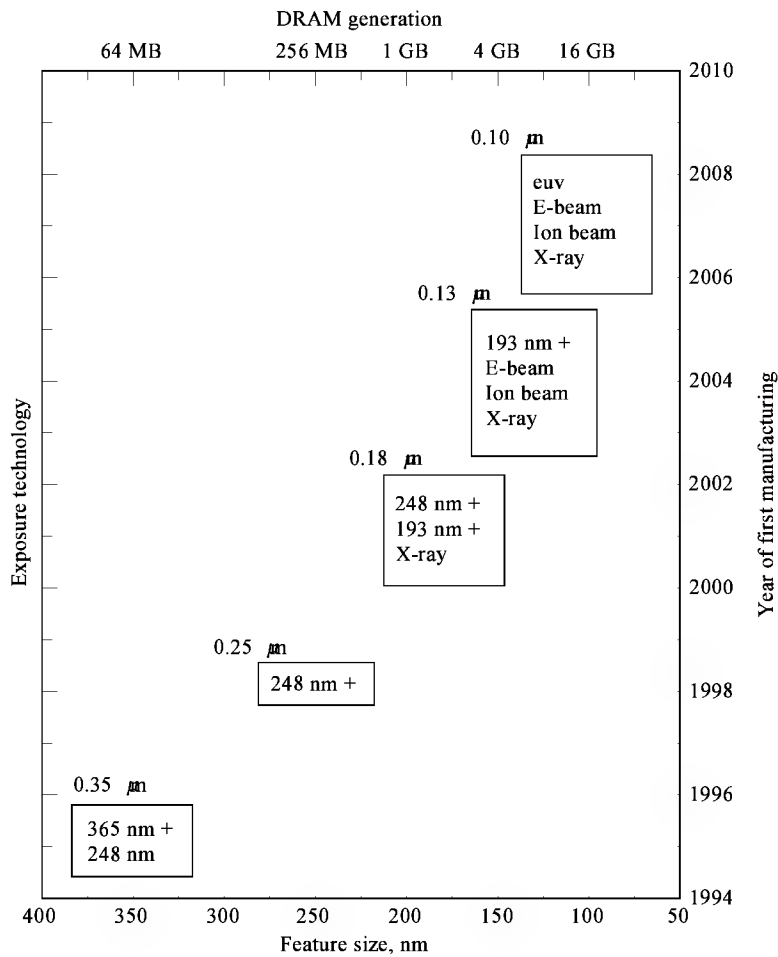


Fig. 34. Lithography approaches under consideration for production of future generations of semiconductor devices. The printed feature size, exposure technology, corresponding generation of dynamic random access memory (DRAM) integrated circuits and the anticipated year of first manufacturing are included. For example, the candidates for manufacturing 4 gigabit DRAM devices at 0.13 μm ground rules (beginning in 2004) are 193-nm lithography with resolution enhancement techniques, or e-beam, ion-beam or x-ray lithography. Optical enhancements or resist modification indicated by + sign.

Figure based on the 1994 SLA roadmap (112).

Thin-Film Imaging Resists. The use of shorter exposure wavelengths and other refinements in exposure tool optics improve resolution; however, this is usually accompanied by an increased difficulty in maintaining the image pattern in focus through the thickness of the resist film; that is, there is a trade-off between exposure tool resolution and its depth-of-focus. A resist material that could be used as a very thin film (at a thickness of ca 200 nm or less), yet still would act as a useful mask throughout subsequent processing steps, would go far in circumventing the optics trade-off. Thin-film imaging (TFI) resists incorporating silicon, when coupled with resist development using oxygen-reactive ion etching (O₂-RIE), are examples of such materials.

Figure 35 shows schematically the process flows of the two primary types of TFI methods, using bilayer and top surface imaged (TSI) resists. Both methods rely on the differences between the oxygen plasma etch rates of polymers that contain elements forming refractory oxides (such as silicon) and simple organic polymers comprised only of elements forming volatile oxides. When a thin film of a resist containing silicon is patterned on top of a polymer underlayer containing no silicon and then placed in an oxygen plasma, the imaged film immediately forms a highly etch-resistant layer of SiO₂ while the unprotected underlayer rapidly erodes. In O₂-RIE the underlayer etch can be performed anisotropically, transferring the pattern of the top film vertically into the thick underlayer. The final bilayer pattern constitutes the resist relief image that is to be used for pattern transfer into the substrate. Since the imaging layer is thin compared to the overall resist thickness, this approach eases the difficulty of maintaining focus, and offers other advantages, which include planarization of underlying substrate topographical features, the potential for reducing substrate reflectivity using opaque underlayers, and the potential for printing images with much higher aspect ratio images than is possible in single-layer resists. These advantages come at the cost of increased process complexity compared to single-layer resist techniques.

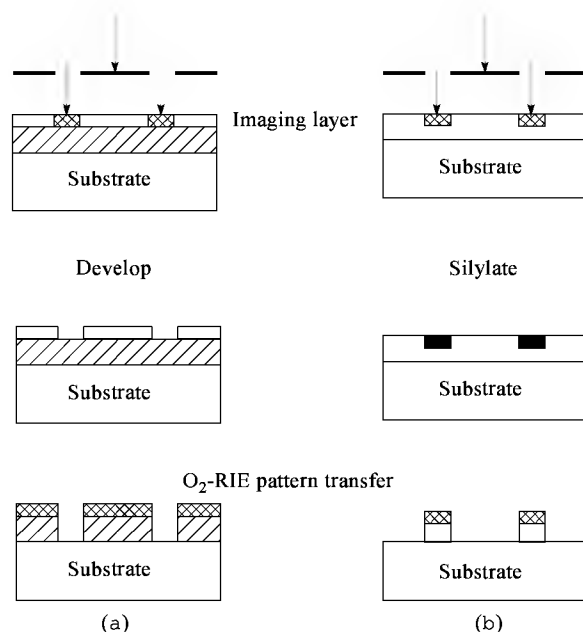


Fig. 35. Process flow for thin-film imaging lithography: (a) bilayer process and (b) top surface imaging. The bilayer process shown here employs a positive-tone imaging layer. The TSI process illustrated relies on preferential silicon incorporation in the exposed regions of the imaging layer to give a negative-tone pattern.

Bilayer Resists. A variety of silicon-containing bilayer resists have been described, with examples of both CA and conventional imaging chemistries in negative and positive tone (115). Resist films containing 10–12% silicon by weight of solids generally show enough O_2 -RIE etch resistance to serve as the imaging layer of a useful bilayer system. The design of such systems is not a simple matter, however, as the introduction of this proportion of silicon (either as an additive, as part of the polymer backbone, or in functional groups pendent to the polymer chain) can strongly perturb the polymer solubility and/or its T_g to the detriment of its lithographic properties. Figure 36 provides several examples of this type of resist.

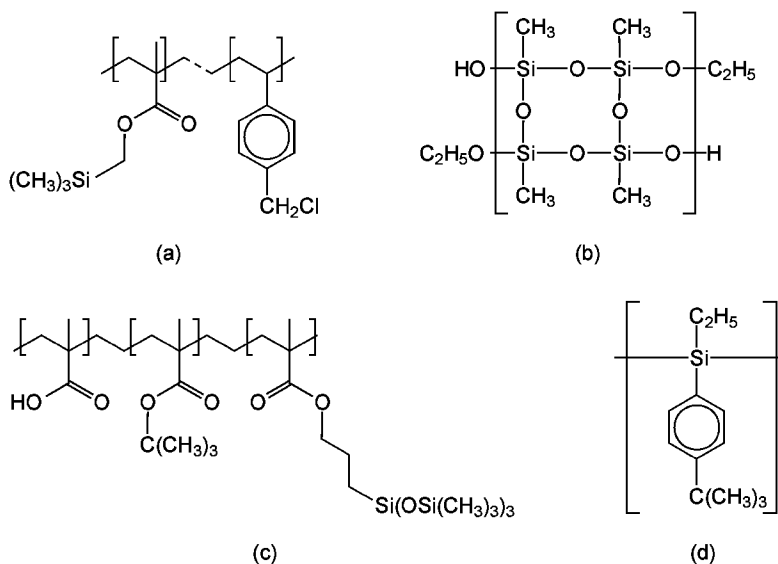


Fig. 36. Representative bilayer resist systems. Both CA and non-CA approaches are illustrated (116–119). (a) Cross-linking E-beam resist, 193-nm thin-film imaging resist; (b) acid-catalyzed negative-tone cross-linking system; (c) positive-tone CA resist designed for 193-nm applications; and (d) positive-tone resist based on chain scission.

Top-Surface Imaging Resists. TSI resists differ from bilayer systems in that the silicon image is introduced into the polymer after the imaging step (120). The same deprotection chemistry used to unmask phenolic hydroxyl groups in CA resists can also be used to spatially direct their silylation when the film is treated with a silylating reagent (121). In principle, this technique should afford very high silylation contrast since the unexposed regions of the film contain no reactive sites. In a second method, radiation-induced polymer cross-linking slows diffusion of a silylation reagent into a polymer film containing functional groups that can combine with the reagent (122). Both CA or non-CA systems of this type have been described. While patterned films show a preferential incorporation of silicon in the uncross-linked areas of the resist, there is measurable silylation in the cross-linked areas. The inherently lower silylation contrast must be overcome by careful optimization of the etch processes and silylation reagent.

In each of these approaches, imaging is confined to the top of a single polymeric film by adjusting optical absorption. The penetration depth of the silylation agent and the attendant swelling of the polymer film must also be controlled to avoid distortion of the silylated image. Resists of this type are capable of very high resolution (Fig. 37).

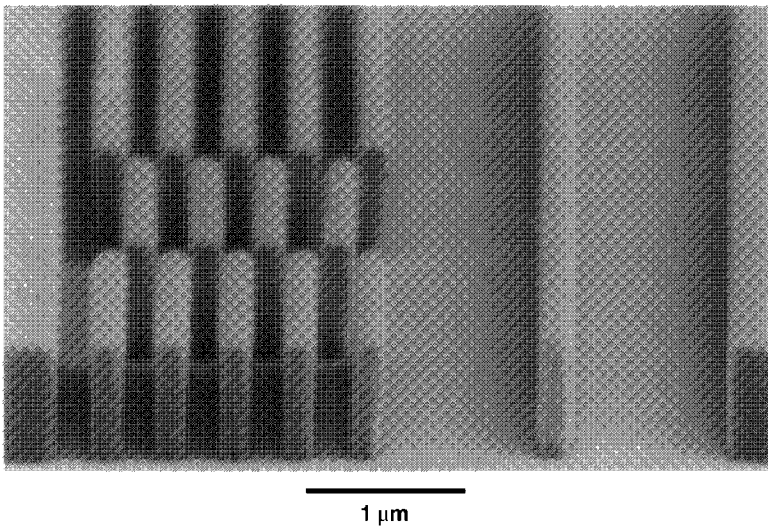


Fig. 37. Resist images obtained with a cross-linking monocomponent TSI resist (PHOST polymer), cross-linked by photo-oxidation using light at 193-nm wavelength. After exposure, the film was treated with a vapor of dimethylsilyldimethylamine and then plasma developed using O₂-RIE (122).

SEM photomicrograph courtesy of MIT Lincoln Laboratory.

Alternative Exposure Technologies.

Optical Enhancements. The term wavefront engineering has been coined to describe a set of optical techniques intended to extend exposure tool resolution beyond the classical Rayleigh limit (123). This can take the form of modifications to the exposure illumination system (such as off-axis illumination) and or modifications to the photomask. In one example, termed phase-shift mask technology, a conventional quartz photomask with transparent regions and opaque chromium features is modified to incorporate optical-phase shifting elements. By controlling both amplitude and phase of the transmitted light, resolution and depth of focus can be improved for specific feature types. Phase-shift mask lithography is best suited to the fabrication of devices comprised of regular arrays such as electronic memory devices. The complex patterns found in logic and processor circuits present a greater challenge in mask fabrication.

Electron-Beam and Ion-Beam Lithography. The serial nature of electron-beam lithography, where only one feature at a time is exposed, severely limits its utility as a full-scale manufacturing technology. Several approaches to increasing throughput are currently under investigation. One such scheme is termed cell projection and makes use of a specially shaped beam to write one or more device cells in a single exposure (124). The beam pattern is generated by projecting the electrons through a shaping aperture in a silicon substrate (Fig. 39). This aperture can have a variety of different cell shapes, including simple square apertures for writing irregular patterns as in conventional e-beam lithography. Proponents of cell-projection technology claim that the partial parallelism inherent in the method, when coupled with high sensitivity CA e-beam resists, should result in a wafer throughput adequate for manufacturing (125). An alternative projection e-beam technique relies on step-and-repeat imaging through a transparent membrane mask at an image reduction ratio of 4×. Pattern delineation is based on differential scattering (rather than absorption) off the respective mask features. The current generation tool is being exercised with CA resists, the resolutions of 0.075 micrometer line-and-space features have been reported (126). In a third proposed enhancement to e-beam lithography, high throughput is attained by effectively operating many miniaturized conventional direct-wire e-beam systems in parallel. This is achieved through the use of microcolumn arrays where each element can act as an individual and independent electron source. Functional columns 3.5 mm long have been constructed, and fabrication of arrays of such devices using silicon micromachining technology is now being pursued (127).

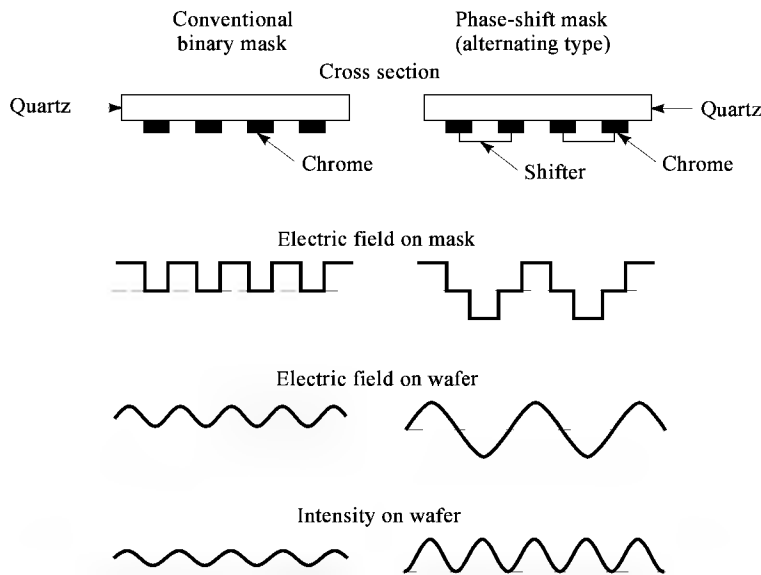


Fig. 38. Diagram comparing the optical characteristics of a standard binary chrome mask with a phase-shift mask. The changes in the electric fields introduced by the phase-shift elements result in a sharper light intensity profile at the wafer surface.

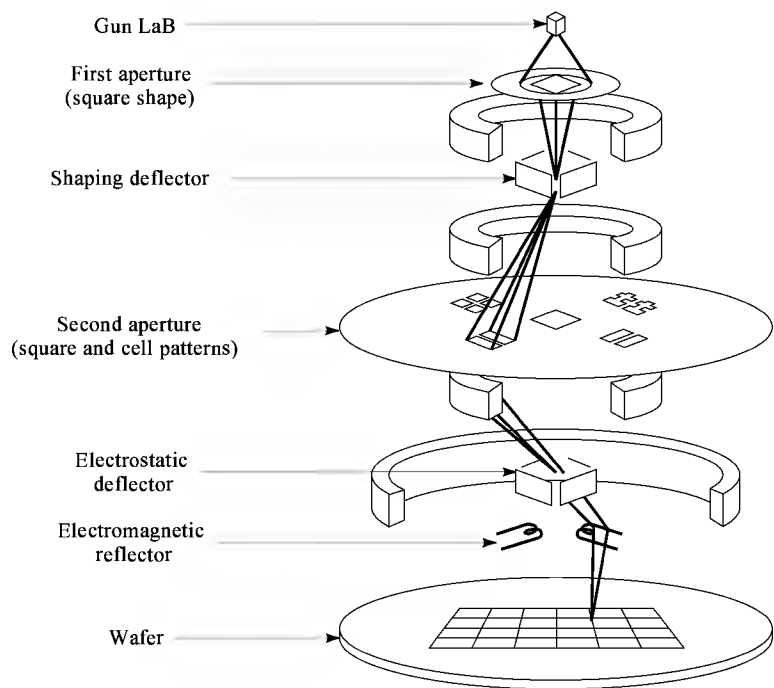


Fig. 39. Schematic showing the basics of cell projection. The desired beam shape is selected by steering the electron beam through the appropriate pattern in the aperture plate. By using a rectangular aperture the system can operate like a conventional direct-write e-beam tool, so any shape of pattern can be created using this approach (125).

Ion-beam lithography is quite similar conceptually to e-beam lithography (128). The larger effective mass of the accelerated ion (typically a proton) and its lower penetration depth into the resist greatly reduces scattering in the resist film compared to e-beam exposure. In consequence the ion-beam analogue has a better intrinsic resolution. Both direct wire and projection ion-beam lithographies are possible, the latter offering higher throughput. Throughput concerns indicate that the use of a sensitive resist is required.

Extreme Uv (Euv) Lithography. This term denotes a projection lithography system designed to use light of wavelength near 13 nm (the soft x-ray region), produced by irradiation of a metal target with a pulsed uv laser (129). At this wavelength, high resolution imaging can be achieved while maintaining acceptable depth of focus if imaging optics with a low numerical aperture are used. However, this is accompanied by a number of technical challenges, including the development of new x-ray sources, an all-reflective optical design and a new type of reflective mask suitable for soft x-ray imaging (Fig. 40). Proponents of euv lithography predict resolution scalable to 0.07 microns while maintaining a large depth of focus.

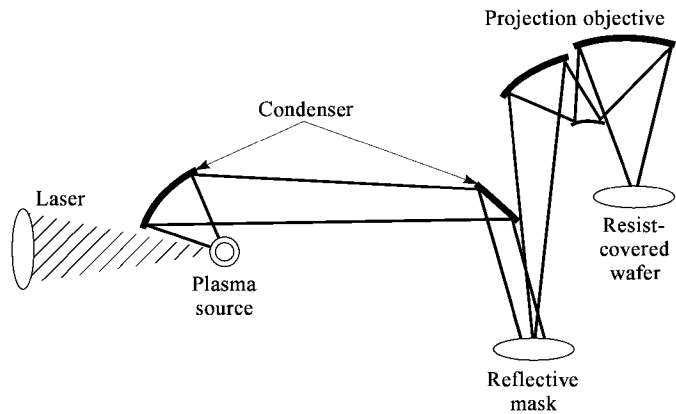


Fig. 40. Schematic of an euv exposure tool. Key features are the excimer laser-driven x-ray source and the reflective optical elements (including the mask) in the imaging system.

Courtesy of Sandia National Laboratories.

Light absorption at a wavelength of 13 nm is determined by the atomic weights of constituent elements of the absorbing material and its density. In general, organic materials strongly absorb euv light, so researchers anticipate that TFI resists will be called for. Given the low brightness of available soft x-ray sources, and the potential for radiation damage to optical components, it is anticipated that high sensitivity CA resists must be used if euv lithography is to be adopted for full scale device manufacturing.

Proximity X-Ray. The practice wherein improved resolution is achieved by decreasing the wavelength of the imaging radiation reaches its limit with x-ray lithography. A significant investment has been made in the use of synchrotrons (Fig. 41) as light sources for lithography, using emission in the spectral range of 1–2 nm (130). The high transparency of organic films at this wavelength, combined with a large effective depth of focus, enable the fabrication of resist structure with high aspect ratios. Proximity x-ray lithography has been used to create features as small as 30 nm (131).

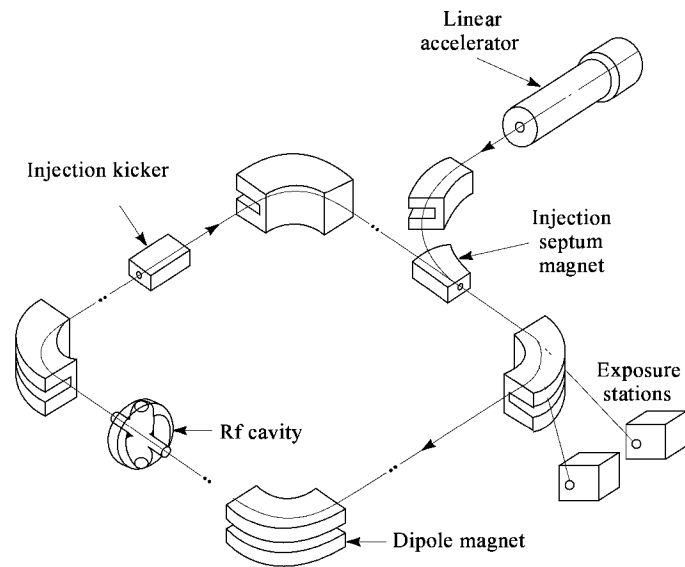


Fig. 41. Diagram of a compact storage ring. Each time the circulating electron beam is bent by one of the dipole magnets it loses energy by emitting x-ray photons which are directed to an exposure station. One ring can serve as the illumination source for many attached exposure stations. A constant ring current is maintained by replacing these energy losses by means of an rf accelerating cavity (130).

The construction of suitable masks is one of the most challenging aspects of x-ray lithography (Fig. 42). The mask substrate must be thin to avoid significant attenuation of the exposing beam but must be sufficiently rugged to survive handling and cleaning. The absorber material must be fabricated as a relatively thick film to achieve adequate blocking of the beam. Finally, since proximity x-ray lithography is a simple shadow patterning technique, it requires a mask with the same dimensions as the ultimate feature to be patterned (optical lithography tools usually project a demagnified image of the mask onto the resist film, simplifying mask fabrication since the mask pattern is four or five times larger than the final resist feature).

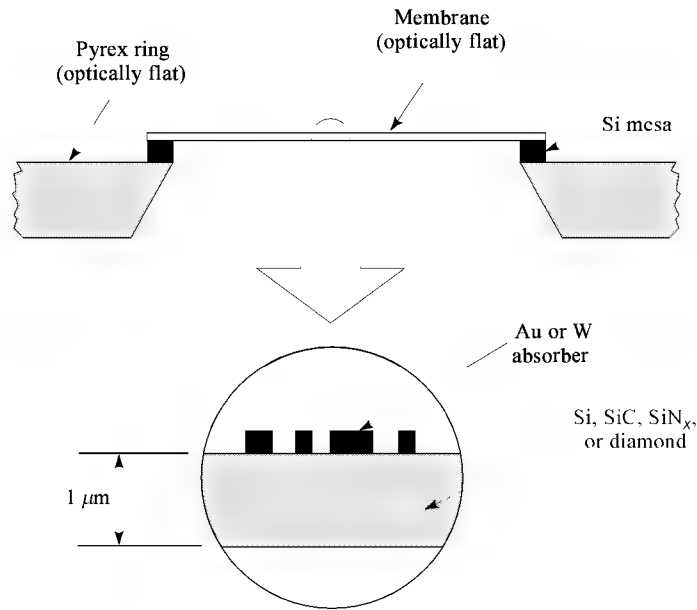


Fig. 42. Diagram of a mask used for 1:1 proximity x-ray lithography. These thin-membrane masks, required for optimum transmission when using patterned metal absorbers, must remain free of distortion to maintain pattern fidelity during exposure.

Even though synchrotrons can provide very high radiation flux, CA resists can be advantageously applied (19). Conventional aqueous-developing resists such as the DNQ–novolac family have been used for very high resolution x-ray lithography, but the radiation sensitivity of such systems is unsuited for manufacturing-scale throughput. The superior radiation sensitivity of CA resists both improves system utilization and reduces mask radiation damage during exposure.

BIBLIOGRAPHY

1. C. G. Willson, L. Thompson and M. Bowden, eds., *Introduction to Microlithography*, 2nd ed., American Chemical Society, Washington, D.C., 1994.
2. W. Moreau, *Semiconductor Lithography*, Plenum Press, New York, 1988.
3. E. Reichmanis and A. Novembre, *Ann. Rev. Mater. Sci.* **23**, 11 (1993).
4. J. Shaw, M. Hatzakis, E. Babich, J. Paraszczak, D. Witman, and K. Stewart, *J. Vac. Sci. Tech. B* **7**(6), 1709 (1989).
5. G. Moore, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 2 (1995).
6. J. Bryzek, K. Pedersen, and W. McCulley, *IEEE Spectrum* **31**(5), 20 (1994).
7. G. Kovacs, K. Petersen, and M. Albin, *Anal. Chem.* **68**, 407A (1996).
8. L. Lin, S. Lee, K. Pister, and M. Wu, *IEEE Photonics Technology Letters* **6**, 1445 (1994) and references therein.
9. D. Cho, R. Warrington Jr., A. Pisano, H. Bau, C. Friedrich, J. Jara-Almonte, and J. Liburdy, eds., *Micromechanical Sensors, Actuators and Systems*, American Society of Mechanical Engineers, New York, 1991.
10. M. Eggers and D. Ehdich, *Hematologic Pathology* **9**(1), 1 (1995).
11. W. DeForest, *Photoresist: Materials and Processes*, McGraw-Hill, New York, 1975.
12. P. Hartsuch, *Chemistry of Lithography*, Lithographic Technical Foundation, New York, 1961.
13. G. Mannivannan, R. Changkakoti, R. Lessard, G. Mailhot, and M. Bolte, *J. Phys. Chem.* **97**, 7228 (1993).

14. L. F. Thompson and R. E. Kerwin, *Ann. Rev. Matls. Sci.* **6** 267 (1976).
15. M. King, in N. Einspruch, eds., *VLSI Electronics: Microstructure Science*, Vol. 1, Academic Press, New York, 1981, Chapt. 2.
16. M. Gurian and N. Ivory, *Electronic Packag. Prod.* **35**(3), 49 (1995).
17. S. Baller, *Printed Circuit Fabrication*, **18**(7), 28 (1995).
18. H. Pfeiffer, *Solid State Technol.* **27**(9), 223 (1984).
19. J. Wadaumont, *J. Vac. Sci. Tech. B* **7**(6), 1634 (1989).
20. J. Lingnau, R. Dammel, and J. Theis, *Solid State Technol.* **32**(9), 105 (1989).
21. J. Lingnau, R. Dammel, and J. Theis, *Solid State Technol.* **32**(10), 107 (1989).
22. L. Thompson and M. Bowden, *J. Electrochem. Soc.* **120**, 1722 (1973).
23. R. Dammel, *Diazonaphthoquinone-based Resists*, SPIE Optical Engineering Press, Bellingham, Wash., 1993.
24. F. Billmeyer, *Textbook of Polymer Science*, 2nd ed., Wiley-Interscience, New York, 1984, pp. 436–439.
25. G. Buhr, H. Lenz, and S. Scheler, *Proc. Soc. Photo-Opt. Instr. Eng.* **1086**, 117 (1989).
26. P. Trefonas and B. Daniels, *Proc. Soc. Photo-Opt. Instr. Eng.* **771**, 194 (1987).
27. C. Szmanda, A. Zampini, D. Madoux, and C. McCants, *Proc. Soc. Photo-Opt. Instr. Eng.* **1086**, 363 (1989).
28. S. Kishimura, A. Yamaguchi, Y. Yamada, and H. Nagata, *Polym. Eng. Sci.* **32**(20), 1550 (1992).
29. K. Uenishi, Y. Kawabe, T. Kokubo, S. Slater, and A. Blakeney, *Proc. Soc. Photo-Opt. Instr. Eng.* **1466**, 102 (1991).
30. K. Honda, B. Beauchemin, R. Hurditch, A. Blakeney, Y. Kawabe, and T. Kobuko, *Proc. Soc. Photo-Opt. Instr. Eng.* **1262**, 493 (1990).
31. J. Stille, *J. Chem. Ed.* **58**(11), 862 (1981).
32. R. Allen, K. Chen, and P. Gallagher-Wetmore, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 250 (1995).
33. M. Hanabata, Y. Uetani, and A. Furuta, *J. Vac. Sci. Tech. B* **7**(4), 640 (1989).
34. M. Hanabata, F. Oi, and A. Furuta, *Proc. Soc. Photo-Opt. Instr. Eng.* **1466**, 132 (1991).
35. M. Lepselter and W. Lynch, in Ref. 15, Chapt. 3.
36. G. Pawlowski, T. Sauer, R. Dammel, D. Gordon, W. Hinsberg, D. McKean, C. Lindley, H. Merrem, H. Roshert, R. Vicari, and G. Willson, *Proc. Soc. Photo-Opt. Instr. Eng.* **1262**, 391 (1990).
37. H. Ito and G. Wilson, *Polym. Eng. Sci.* **23**, 1012 (1982).
38. J. Frechet, H. Ito, and C. G. Willson, *Proc. Microcircuit Eng.*, **82**, 260 (1982).
39. D. McKean, U. Schaedli, and S. MacDonald, in E. Reichmanis, S. MacDonald and T. Iwayanagi, eds., *Polymers in Microlithography, ACS Symposium Series 412*, American Chemical Society, Washington, D.C., 1989, pp. 27–38.
40. J. Maltabes, S. Holmes, J. Morrow, R. Barr, M. Hakey, G. Reynolds, W. Brunsvold, G. Willson, N. Clecak, S. MacDonald, and H. Ito, *Proc. Soc. Photo-Opt. Instr. Eng.* **1262**, 2 (1990).
41. E. Reichmanis, F. Houlihan, O. Nalamasu, and T. Neenan, *Chem. Matls.* **3**, 394 (1991).
42. J. Crivello, in J. P. Fouassier and J. F. Rabek, eds., *Radiation Curing in Polymer Science and Technology (Vol II): Photoinitiating Systems*, Elsevier, New York, 1993, Chapt. 8, for a review on onium salts.
43. H. Ito and G. Willson, in T. Davidson, ed., *Polymers in Electronics, ACS Symposium Series 242*, American Chemical Society, Washington, D.C., 1984, pp. 11–24.
44. R. Allen, G. Wallraff, W. Hinsberg, L. Simpson, and R. Kunz, in *Polymers for Microelectronics, ACS Symposium Series 537*, American Chemical Society, Washington, D.C., 1994, p. 165.
45. K. Nakano, K. Maeda, S. Iwasa, T. Ohfuji, and E. Hasegawa, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 433 (1995).
46. F. Saeva, D. Breslin and H. Luss, *J. Amer. Chem. Soc.* **113**, 5333 (1991).
47. M. F. Cronin, T. Adams, T. Fedynyshyn, J. Georger, J. Michael Mori, R. Sinta, and J. W. Thackeray, *Proc. Soc. Photo-Opt. Instr. Eng.* **2195**, 214 (1994).
48. J. March, *Advanced Organic Chemistry*, 4th ed., Wiley-Interscience, New York, 1992, p. 250.
49. T. Itani, H. Yoshino, M. Fujimoto, and K. Kasama, *J. Vac. Sci. Technol. B* **13**, 2026 (1995).
50. D. McKean, R. Allen, P. Kasai, U. Schaedli, and S. MacDonald, *Proc. Soc. Photo-Opt. Instr. Eng.* **1672**, 94 (1992).
51. H. Ito, *Proc. Soc. Photo-Opt. Instr. Eng.* **920**, 33 (1988).
52. G. Wallraff, R. Allen, W. Hinsberg, L. Simpson, and R. Kunz, *Chemtech* **23**(4), 22 (1993).
53. N. Hacker, in J. P. Fouassier and J. F. Rabek, eds., *Radiation Curing in Polymer Science and Technology*, Volume II, *Photoinitiating Systems*, Elsevier, New York, 1993, Chapt. 9.
54. R. DeVoe, M. Sahyun, and E. Schmidt, *Can. J. Chem.* **66**, 319 (1988).
55. R. Binkley and T. Flechtner in W. Horspool, ed., *Synthetic Organic Photochemistry*, Plenum Press, New York, 1984, Chapt. 7.
56. T. Neenan, F. Houlihan, E. Reichmanis, J. Kometani, B. Bachman, and L. Thompson, *Macromolecules* **23**, 145 (1990).
57. N. Hayashi, L. Schlegel, T. Ueno, H. Shiraishi, and T. Iwayanagi, *Proc. Soc. Photo-Opt. Instr. Eng.* **1466**, 377 (1991).
58. L. Schlegel, T. Ueno, H. Shiraishi, N. Hayashi, and T. Iwayanagi, *Chem. Mater.* **2**, 299 (1990).
59. K. Naitoh, K. Yoneyma, and T. Yamaoka, *J. Phys. Chem.* **96**, 238 (1992).
60. H. Ito, G. Breyta, D. Hofer, T. Fischer, and B. Prime, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 53 (1995).
61. M. Shirai and M. Tsunooka, in *Progress in Polymer Science*, Vol. 21, Elsevier, UK, 1996, pp. 1–45.
62. T. H. Lowery and K. S. Richardson, *Mechanism and Theory in Organic Chemistry*, 3rd ed., Harper and Row, New York, 1987, Chapt. 7.
63. Ref. 48, Table 10.14, p. 380.
64. D. McKean, S. MacDonald, N. Clecak, and G. Willson, *Proc. Soc. Photo-Opt. Instr. Eng.*, **1925**, 246 (1993).
65. R. Allen, Q. Ly, G. Wallraff, C. Larson, W. Hinsberg, W. Conley, and K. Muller, *Proc. Soc. Photo-Opt. Instr. Eng.* **1925**, 246 (1993).
66. H. Ito, M. Ueda, and R. Schwalm, *J. Vac. Sci. Technol. B* **6**, 2259 (1988).
67. D. Funhoff, H. Binder, and R. Schwalm, *Proc. Soc. Photo-Opt. Instr. Eng.* **1672**, 46 (1992).
68. G. Wallraff, J. Hutchinson, W. Hinsberg, F. Houle, P. Seidel, R. Johnson, and W. Oldham, *J. Vac. Sci. Technol. B* **12**, 3857 (1994).
69. G. M. Wallraff, J. Hutchinson, W. Hinsberg, F. Houle, and P. Seidel, *Microelectronic Eng.* **27**, 397 (1995).
70. W. Huang, R. Kwong, A. Katani, and M. Khojasteh, *Proc. Soc. Photo-Opt. Instr. Eng.* **2195**, 37 (1994).
71. K. Przybilla, Y. Kinoshita, T. Kudo, S. Masuda, H. Okazaki, M. Padmanaban, G. Pawlowski, H. Roeschert, W. Speiss, and N. Suehiro, *Proc. Soc. Photo-Opt. Instr. Eng.* **1925**, 76 (1993).
72. T. Hattori, A. Imai, R. Yamanaka, T. Ueno, and H. Shiraishi, *J. Photopolym. Sci. Technol.* **9**, 611 (1996).
73. R. Allen, W. Conley, and J. Gelorme, *Proc. Soc. Photo-Opt. Instr. Eng.* **1672**, 513 (1992).
74. W. Feeley, J. Imhof, and C. Stein, *Polym. Eng. Sci.* **26**, 1101 (1986).
75. J. Frechet, S. Matuszczak, H. Stover, C. Willson, and B. Reck, in E. Reichmanis, S. A. MacDonald and T. Iwayanagi, eds., *Polymers in Microlithography, ACS Symposium Series 412*, American Chemical Society, Washington, D.C., 1989, p. 74.
76. S. MacDonald, N. Clecak, H. Wendt, G. Willson, C. Snyder, C. Knors, N. Deyoe, J. Maltabes, J. Morrow, A. McGuire, and S. Holmes, *Proc. Soc. Photo-Opt. Instr. Eng.* **1466**, 1 (1991).
77. O. Nalamasu, E. Reichmanis, M. Cheng, V. Pol, J. Kometani, F. Houlihan, T. Neenan, M. Bohrer, D. Mixon, and L. Thompson, *Proc. Soc. Photo-Opt. Instr. Eng.* **1466**, 13 (1991).
78. R. Cox, L. Druet, A. Klausner, T. Modro, P. Wan, and K. Yates, *Can. J. Chem.* **59**, 1568 (1981).
79. A. Oikawa, Y. Hatakenaka, Y. Ikeda, Y. Kokubo, M. Tanishima, N. Sanoth, and N. Abe, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 599 (1995).
80. T. Kumeda, Y. Tanaka, A. Ueyama, S. Kubota, H. Koezuka, T. Hanawa, and H. Morimoto, *Proc. Soc. Photo-Opt. Instr. Eng.* **1925**, 31 (1993).

81. A. Oikawa, N. Santoh, S. Miyata, and Y. Hatakenaka, *Proc. Soc. Photo-Opt. Instr. Eng.* **1925**, 92 (1993).

82. Y. Kawai, A. Otaka, A. Tanaka, and T. Matsuda, *Jpn. J. Appl. Phys.* **33**, 7023 (1994).

83. Y. Kawai, A. Otaka, J. Nakamura, A. Tanaka, and T. Matsuda, *J. Photopolym. Sci. Techn.* **8**, 535 (1995).

84. K. Asakawa, T. Ushirogouchi, and M. Nakase, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 563 (1995).

85. S. Satio, N. Kihara, T. Naito, M. Nakase, T. Nakasugi, and Y. Kato, *J. Photopolym. Sci. Technol.* **9**, 677 (1996).

86. J. E. McGrath, *J. Chem. Ed.* **58**(11), 844 (1981).

87. R. Dammel, M. Rahman, P. Lu, A. Canize, and V. Elango, *Proc. Soc. Photo-Opt. Instr. Eng.* **2195**, 542 (1994).

88. O. Nalamasu, A. Timko, M. Cheng, J. Kometani, M. Galvin, S. Heffner, S. Slater, A. Blakeney, N. Munzel, R. Schulz, H. Holzwarth, C. Mertesdorf, and T. Schacht, *Proc. Soc. Photo-Opt. Instr. Eng.* **1925**, 155 (1993).

89. H. Ito, G. Breyta, D. Hofer, R. Sooriyakumaran, K. Petrillo, and D. Seeger, *J. Photopolym. Sci. Tech.* **7**, 433 (1994).

90. R. Allen, G. Wallraff, W. Hinsberg and L. Simpson, *J. Vac. Sci. Technol. B* **9**, 3357 (1991).

91. Y. Kaimoto, K. Nozaki, S. Takechi, and N. Abe, *Proc. Soc. Photo-Opt. Instr. Eng.* **1672**, 66 (1992).

92. R. Allen, G. Wallraff, R. DiPietro, D. Hofer, and R. Kunz, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 474 (1995).

93. K. Nozaki, K. Watanabe, E. Yano, A. Kotachi, S. Takechi, and I. Hanyu, *J. Photopolym. Sci. Technol.* **9**(3), 509 (1996).

94. N. Shida, T. Ushirogouchi, K. Asakawa, and N. Nakase, *J. Photopolym. Sci. Technol.* **9**(3), 457 (1996).

95. R. Allen, R. Sooriyakumaran, J. Opitz, G. Wallraff, R. DiPietro, G. Breyta, D. Hofer, R. Kunz, S. Jayaraman, R. Schick, B. Goodal, U. Okoroanyanwu, and G. Willson, *Proc. Soc. Photo-Opt. Instr. Eng.* **2724**, 334 (1996).

96. T. Wallow, F. Houlihan, O. Nalamasu, E. Chandross, T. Neenan, and E. Reichmanis, *Proc. Soc. Photo-Opt. Instr. Eng.* **2724**, 355 (1996).

97. T. Ohfuju, O. Nalamasu, and D. Stone, *J. Vac. Sci. Technol. B* **11**, 2714 (1993).

98. W. Conley, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 41 (1995).

99. J. Much, D. Hess, and E. Aydil, in ref. 1, pp. 377–493.

100. L. Pedersen, *J. Electrochem. Soc.* **129**, 205 (1982).

101. H. Gokan, S. Esho, and Y. Ohnishi, *J. Electrochem. Soc.* **130**(1), 143 (1983).

102. R. Kunz, S. Palmateer, A. Forte, R. Allen, G. Wallraff, R. DiPietro, and D. Hofer, *Proc. Soc. Photo-Opt. Instr. Eng.* **2724**, 365 (1996).

103. W. Hinsberg, S. MacDonald, C. Snyder, H. Ito, and R. Allen, in *Polymers for Microelectronics, ACS Symposium Series 537*, American Chemical Society, Washington, D.C., 1994, p. 101.

104. L. Schlegel, T. Ueno, N. Hayashi, and T. Iwayanagi, *J. Vac. Sci. Technol. B* **9**, 278 (1991).

105. F. Vinet, N. Buffet, P. Fanton, L. Pain, and P. Paniez, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 202 (1995).

106. W. Hinsberg, S. MacDonald, N. Clecak, and C. Snyder, *Proc. Soc. Photo-Opt. Instr. Eng.* **1672**, 24 (1992).

107. W. Hinsberg, S. MacDonald, N. Clecak, C. Snyder, and H. Ito, *Proc. Soc. Photo-Opt. Instr. Eng.* **1925**, 43 (1993).

108. P. Paniez, C. Rosilio, B. Mouanda, and F. Vinet, *Proc. Soc. Photo-Opt. Instr. Eng.* **2195**, 14 (1994).

109. H. Ito, W. England, N. Clecak, G. Breyta, H. Lee, D. Yoon, R. Sooriyakumaran, and W. Hinsberg, *Proc. Soc. Photo-Opt. Instr. Eng.* **1925**, 65 (1993).

110. G. Breyta, D. Hofer, H. Ito, D. Seeger, K. Petrillo, H. Moritz, and T. Fischer, *J. Photopolym. Sci. Tech.* **7**, 449 (1994).

111. *Proceedings of the Society of Photo-Optical Instrumentation Engineers, Journal of Vacuum Science and Technology, Journal of Electrochemical Society, Micro/NanoEngineering, and Chemistry of Materials*; these are among those in which new work in advanced resist chemistry can be located.

112. K. H. Brown, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 33 (1995). A good source of information on the new exposure technologies can be found in the Proceedings of the International Conference on Electron, Ion and Photon Beam Technology and Nanofabrication, published annually in Issue 6 of the *Journal of Vacuum Science and Technology B*.

113. R. DeJule, *Semiconductor Int.* **19**(9), 85 (1996).

114. P. N. Dunn, *Solid State Technol.* **37**(6), 49 (1994).

115. R. D. Miller and G. M. Wallraff, *Adv. Mater. Optics Electron.* **4**, 95 (1994).

116. B. Smith, D. Mixon, A. Novembre, and S. Butt, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 504 (1995).

117. H. Watanabe, Y. Todokoro, and M. Inoue, *J. Vac. Sci. Technol. B* **9**, 3436 (1991).

118. U. Schaedeli, E. Inguely, A. Blakeney, P. Falcigno, and R. Kunz, *Proc. Soc. Photo-Opt. Instr. Eng.* **2724**, 344 (1996).

119. G. Wallraff, R. Miller, M. Baier, E. Ginsberg, and R. Kunz, *J. Photopolym. Sci. Technol.* **5**, 111 (1992).

120. D. Seeger, D. La Tulipe, R. Kunz, C. Garza, and M. Hanratty, *IBM J. Res. Devel.* **41**, 105 (1997).

121. S. MacDonald, G. Willson, and J. Frechet, *Acc. Chem. Res.* **27**, 151 (1994).

122. S. Palmateer, R. Kunz, M. Horn, A. Foote, and M. Rothschild, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 455 (1995).

123. M. D. Levenson, *Solid State Technol.* **38**(2), 57 (1995).

124. Y. Nakayama, S. Okazaki, N. Saitou, and H. Wakabayashi, *J. Vac. Sci. Technol. B* **8**, 1836 (1990).

125. S. Okazaki, *Proc. Soc. Photo-Opt. Instr. Eng.* **2438**, 18 (1995).

126. L. R. Harriott, S. Berger, C. Biddick, M. Blakey, S. Bowler, K. Brady, R. Camarda, W. Connelly, A. Crocken, J. Custy, R. Dimarco, R. Farrow, J. Felker, L. Fetter, R. Freeman, L. Hopkins, H. Huggins, C. Knurek, J. Kraus, J. Liddle, M. Mkrtychan, A. Novembre, M. Peabody, R. Tarascon, H. Wade, W. Waskiewicz, G. Watson, K. Werder, and D. Windt, *J. Vac. Sci. Technol. B* **14**, 3825 (1996).

127. T. Chang, M. Thompson, E. Kratchmer, H. Kim, M. Yu, Y. Lee, S. Rishton, B. Hussey, and S. Zolgharnain, *J. Vac. Sci. Technol. B* **14**, 3774 (1996).

128. H. Loschner, G. Stengl, A. Chalupka, J. Fegerl, R. Fischer, G. Lammer, L. Malek, R. Nowak, C. Traher, and P. Wolf, *J. Vac. Sci. Technol. B* **11**, 487 (1993).

129. G. Kubiak and D. Kania, eds., *OSA Trends in Optics and Photonics on Extreme Ultraviolet Lithography*, Optical Society of America, Washington, D.C., 1996.

130. M. Wilson, A. Smith, V. Kempson, M. Townsend, J. Schouten, R. Anderson, A. Jorden, V. Suller, and M. Poole, *IBM J. Res. Develop.* **37**, 357 (1993).

131. H. Smith and M. Schattenberg, *IBM J. Res. Develop.* **37**, 319 (1993).

W. D. Hinsberg
G. M. Wallraff
R. D. Allen
IBM Research Division

MICROBIAL AND VIRAL FILTRATION

Several physicochemical methods exist to ensure the safety of biological and biopharmaceutical products in manufacture. Whereas physical methods such as heat and radiation may be used, these are often not viable options owing to detrimental effects on product quality. For example, in the case of products that are heat labile or biochemically complex, these methods may result in alteration of the chemistry or function of the product, or random adsorption of key components. Filtration (qv) is the separation of particles from a fluid (liquid or gas) by passage of that fluid through a permeable medium. Sterile filtration ensures complete removal of viable organisms. Advances in membrane technology (qv) have resulted in the availability of filtrative devices for the removal of viruses in addition to bacteria. Thus membrane filtration is becoming increasingly the method of choice for sterilization of biologicals, especially when the product is heat labile, because the filtration process is inherently nondestructive.

An overview of the general principles of filtration having specific application to bacterial and viral removal is given herein. The emphasis is on ensuring that the sterility and or safety of biologicals and biopharmaceuticals be maintained.

Filtration for Bacterial Removal

The introduction of parenteral drugs at the beginning of the twentieth century created a concomitant need for a suitable method to ensure adequate sterilization of these biochemically complex heat-labile products. Several different types of filters were introduced into the industrial arena: porcelain filter cartridges, asbestos–cellulose layers, and membrane filters. The porcelain filters (Chamberlain) were used extensively for the manufacture of antibiotics. Problems associated with cleaning and concerns over cross-contamination of products arose, however, and these filters fell into disuse. The first filter medium to be used on an industrial scale was the Seitz EK Filter (EK *entkeimung* or germ removal). The asbestos content of that filter limited its applications. The first membrane disks were introduced in 1929 and represented a breakthrough in filtration technology. Four decades later membrane filter cartridges were introduced for processing large batches of parenterals.

The earliest commercially available filters were manufactured in two pore sizes: 0.45 and 0.8 μm. The 0.45 μm-rated membranes were considered to be sterilizing-grade filters and were successfully used in the sterile filtration of pharmaceuticals and parenterals. The membrane filters were qualified using *Serratia marcescens*, a standard bacterium, having dimensions of 0.6 × 1 μm. However, in the late 1960s it became apparent that the matrix of the 0.45 μm-rated filters could be penetrated by some pseudomonad-like organisms (1). For sterile filtration applications in the 1990s, 0.2 μm-rated membranes are the industry standard in the manufacture of sterile parenterals and pharmaceuticals.

General Principles.

Mechanisms of Filter Retention. In general, filtrative processes operate via three mechanisms: inertial impaction, diffusional interception, and direct interception (2). Whereas these mechanisms operate concomitantly, the relative importance and role of each may vary.

Direct interception refers to a sieve-type mechanism in which contaminants larger than the filter pore size are directly trapped by the filter. This sieve retention mechanism of particle arrest is the mechanism of choice and occurs owing to geometric or spatial restraint. This type of particle arrest is considered to be absolute, that is, it is independent of filtration conditions.

Inertial impaction involves the removal of contaminants smaller than the pore size. Particles are impacted on the filter through inertia. In practice, because the differential densities of the particles and the fluids are very small, inertial impaction plays a relatively small role in liquid filtration, but can play a major role in gas filtration.

Diffusional interception or Brownian motion, ie, the movement of particles resulting from molecular collisions, increases the probability of particles impacting the filter surface. Diffusional interception also plays a minor role in liquid filtration. The nature of liquid flow is to reduce lateral movement of particles away from the fluid flow lines.

Types of Filters. In general, there are two types of filters used for microbial removal: depth and membrane. The first type removes microorganisms and particles mainly through retention by entrapment or impaction and adherence. These rely on filter matrix depth to achieve particulate contaminant retention. However, using depth filters, microbial cells may sometimes be set free because of high differential pressure, the high flux of fluid passing through the filter, or filter matrix shifting. The primary mechanism of bacterial cell retention by membrane filters is the sieving effect, due to the highly stable, uniform pore matrix, so that trapped cells are not released. Thus membrane filtration is the method of choice for pharmaceutical and biological applications where absolute microbe retention is required. Membrane filters used for sterile filtration applications are typically constructed from, but not limited to, polymers such as cellulose esters, nylon, polyesters, polytetrafluoroethylene (PTFE), poly(vinylidene fluoride) (PVDF), polycarbonate, polypropylene, and polysulfone.

Membrane Filter Ratings. Filters are rated based on the ability to remove particles of a specific size from a fluid. There is, however, no standard on which method is to be used to specify performance. In general, the absolute rating, or cutoff point of a filter refers to the diameter of the largest particle, normally expressed in micrometers which can pass through the filter. The absolute rating is determined under carefully controlled conditions using industry-accepted reference standards, such as silica suspensions, latex beads, or microorganisms. In a filtration system, the actual form of the contaminants is not necessarily spherical, in which case the nominal diameter is generally taken as the largest of the linear dimensions. Many filter manufacturers use a nominal filter rating, which is an arbitrary value determined by the filter manufacturer and expressed in terms of percentage retention of a specific test contaminant (usually latex or glass beads) of a given size distribution. This nominal rating also represents a nominal efficiency figure (~50–95%), or more correctly a degree of filtration. Nominal rating standards are, however, arbitrary and thus comparison of nominally rated filters is imprecise. This rating system is not used to characterize sterilizing-grade filters.

Sterilizing-grade filters require that biological retention capability be evaluated using a microbial challenge test. *Brevundimonas (Pseudomonas) diminuta* (ATCC 19146) is the standard test microorganism used for the validation of sterilizing-grade filters. By U.S. FDA definition, sterilizing-grade filters refer to filters which can remove a minimum concentration of 10⁷ colony forming units per square centimeter (cfu cm²) of *Brevundimonas (Pseudomonas) diminuta* (ATTC 19146) and yield sterile effluent (3).

Filter Selection. A variety of product- and process-related factors govern filter selection. Considerations include the characteristics of the fluid to be filtered, ie, its chemical composition and compatibility with the filtration system (inclusive of the membrane, filter hardware, piping, etc), the level of bioburden present, specifications on effluent quality, the volume of product to be filtered, flow rate, and temperature.

Membrane-Feed Compatibility. The feed stream must be compatible with the membrane selected. The composition of the feed as well as pH and operating temperature must be considered. Materials having excessively low or high pH may not be compatible with certain membrane polymers. The temperature of operation must fall within the membrane manufacturer's recommended temperature range. Most fluids that are sterile filtered are water-based and thus compatible with most membrane materials, such as polyamides, PVDF, polysulfone, and cellulose acetate. In addition to the compatibility of the membrane itself, the compatibility of the filter components, such as the cage, core, and supporting materials, must also be considered. For sterile processes, the biological safety of the membrane filter or filter cartridge must be demonstrated by the performance of the USP <88> Class VI (121 C) Plastics Test for Biological Reactivity. Because there is no specific listing for materials in contact with pharmaceutical products, filter materials of construction are often selected based on the listing for food contact in the *Code of Federal Regulations* (4).

Effluent Quality. The criteria to be met by the effluent or filtrate must be clearly defined. For aseptic processes a typical requirement is sterilization through

an 0.2 μm -rated sterilizing-grade filter, as defined by ASTM Standard F838-83 (5). The objective of the filtration is to remove contaminants from the process feed without compromising product integrity. Consequently, the filter system must not add anything significant, eg, filter materials of construction or constituents extracted therefrom, to the process fluid, neither should a desired component be removed from the fluid being filtered. Sources that could possibly contribute contaminants to the process fluid include the piping and connections in a process, the filter or associated components such as the cage, core, and supporting materials, or the membrane itself. In some instances, there may be what is referred to as migration of media, ie, dislodging and sloughing of the membrane or support material. Many filter manufacturers document the extractables level of a particular filter in an appropriate solvent in terms of a nonvolatile residue (NVR). These extractables are typically composed of the oligomers of materials present in the filter element or additives. Filter manufacturers address the issue of effluent quality requirements by the performance of appropriate tests on filter samples from manufacturing lots. These tests include cleanliness: per USP limits under Particulate Matter in Injections <788> and conformance with requirements for a nonfiber-releasing filter (6,7); oxidizable substances: per USP limits under Purified Water after Flushing; pH: per USP limits under Purified Water after Flushing; and pyrogens: per USP limits under Bacterial Endotoxins Test as determined using the Limulus amoebocyte lysate (LAL) reagent with an aliquot from a soak solution.

Flow Rate Requirements. Flow rate, measured in units of volume per unit time, is dependent on pressure, P , and resistance, R . The flow rate achievable through a filtration system is directly related to the applied differential pressure and inversely related to the resistance to flow. Viscosity of a fluid has a direct effect on resistance. Fluids having higher viscosities are more resistant to flow and the resultant pressure drop across the filter is greater. Consequently, a greater applied pressure is required to maintain the process flow rate. The smaller the pore size rating of a filter, for a fluid of a given viscosity, the greater the resistance to flow. All other factors being equal, if the pressure on a fluid is increased, then the flow rate of that fluid increases. However, if the resistance to flow is increased, such as by membrane plugging, then flow decreases. Additionally, an initial high pressure (and flow rate) can lead to premature plugging, especially for products that may contain gels, as, for example, biological products.

Pressure Considerations. All components of a system contribute a resistance to flow which results in pressure drop. Pressure drops or losses in a system can be caused by piping, connections, valves, and filling heads, as well as by the filter and its assembly. The pressure drop through a clean filter assembly results from the sum of component pressure drops, including filter housing, filter hardware, and filter membrane. As filtration progresses and particulates clog the filter, pressure drop increases. In choosing a filter, therefore, it is necessary to provide an adequate pressure source, not only to overcome the resistance of the filter, but also to permit flow to continue at an acceptable rate as the medium plugs. This allows full use of the effective dirt-holding capacity of the filter. If the ratio of the initial clean pressure drop through the filter to total available pressure is disproportionately high, unacceptable flow quickly results even though the capacity of the membrane for collecting dirt has not been exhausted.

Temperature Requirements of a Process. The temperature of filtration may affect the viscosity of the fluid, the corrosion rate of the housing, and filter medium compatibility. Elevated temperatures tend to accelerate corrosion and may weaken the gaskets and seals of filter housings. In general, the viscosity of fluids decreases with increasing temperature. Filtration of highly viscous fluids may be conducted at elevated temperature (8). For example, pharmaceutical products containing oil or a lipid emulsion as a drug carrier may require filtration at elevated temperature to enhance filterability characteristics.

Sterile Filtration of Liquids. The only true test of a sterilizing-grade filter is its microbial retention capability. By FDA definition, sterilizing-grade 0.2 μm -rated filters refer to filters which can remove more than 10^7 cfu cm^2 of *B. diminuta* and yield sterile effluent. The challenge method for conducting a liquid bacterial challenge is detailed by ASTM F838-83 (5). A schematic of a test setup is presented in Figure 1. The bacterial challenge, *B. diminuta* at a minimum concentration of 1×10^7 cfu cm^2 of filter area, is passed through the filter suspended in a sterile carrier fluid under standard test conditions. Bacterial concentrations are determined in the input as well as in the effluent. Standard microbiological methods on standard cultivation media, such as Mueller Hinton agar or tryptic soy agar, are used. The entire effluent from the test filter, as well as aliquots of dilutions of effluent, are passed through an analysis membrane. Post-challenge recoveries done using analysis membranes allow for assay of the entire effluent so that even a single microorganism in the effluent is detected.

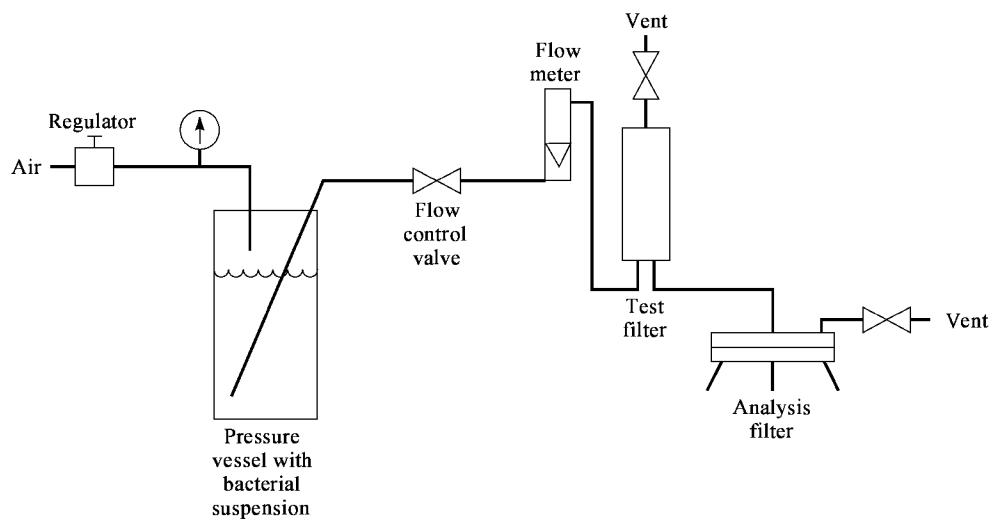


Fig. 1. Schematic of an experimental liquid bacterial setup for evaluation of bacterial retention.

Occasionally, other bioburden organisms may be substituted as the challenge organism. *B. diminuta* is an ideal challenge organism for several reasons. It was a process isolate recovered from effluents (1) following filtration through 0.45 μm -rated filters, which were at one time considered sterilizing-grade filters. *B. diminuta* is an aerobic, asporogenous (does not form spores), gram-negative bacillus that is roughly $0.3 \times 0.8 \mu\text{m}$ in size, and when grown under limiting growth conditions exists in a single cell form.

Important variables in a microbial retention test are (1) culturing conditions of the microorganisms for use as challenge: culturing conditions affect size, shape, and aggregation state of the bacteria. (2) the challenge load or bioburden: the specification for challenge level, a minimum of 1×10^7 cfu cm^2 of filter area, is far in excess of the bioburden routinely encountered in typical pharmaceutical process streams. Challenge concentrations higher than 1×10^8 cfu cm^2 are not recommended as caking/plugging from excessive bioburden can occur; (3) process considerations during testing: passage of microorganisms through partially retentive filters may be enhanced by application of high pressure; temperature may affect fluid viscosity as well as viability and growth rates of microorganisms; and (4) solution characteristics of the carrier fluid: the pH, ionic strength, osmolarity, and presence of additives such as surfactants, especially in the case of simulated process streams, may influence microbial retention; it is necessary to ensure that these variables are controlled in the test, to assure reproducibility of test results.

Factors that could potentially affect microbial retention include filter type, eg, structure, base polymer, surface modification chemistry, pore size distribution, and thickness; fluid components, eg, formulation, surfactants, and additives; sterilization conditions, eg, temperature, pressure, and time; fluid properties, eg, pH, viscosity, osmolarity, and ionic strength; and process conditions, eg, temperature, pressure differential, flow rate, and time.

The efficiency of the filter is evaluated in terms of the titer reduction or log reduction value (LRV). The titer reduction is the ratio of the number of microorganisms in the input suspension to the number of organisms in the effluent. Similarly, the LRV is the $\log_{10}$ of this ratio. The ratio of the difference between the numbers of challenge microorganisms recovered upstream and downstream of the test filter to the average total challenge received by the filter

provides an indication of the removal efficiency of the filter, ie,

$$\text{removal efficiency, \%} = \frac{\text{average total challenge} - \text{average total recovery}}{\text{average total challenge}} \times 100$$

When the filtrate is sterile, the number 1 is substituted for the average total recovery and the value is expressed as greater than the calculated value.

Although *Brevundimonas (Pseudomonas) diminuta* (ATCC 19146) is most commonly used for sterilizing-grade filter validation, in certain applications other bacteria are used. For example, when it is necessary to demonstrate removal of mycoplasma in applications involving sera and tissue culture media, membranes having a smaller pore size rating, eg, 0.1 µm, are frequently used. For these membranes, *Acholeplasma laidlawii* may be employed for validation purposes (9).

Integrity Testing. The only test of a filter's ability to remove bacteria is demonstration of its performance characteristics by bacterial-retention testing. However, a bacterial challenge is a destructive test and precludes subsequent use in a filtration operation. Therefore, filter manufacturers provide validation documentation for a filter with correlation of microbial removal to other nondestructive physical integrity tests. Examples of nondestructive tests most commonly used in the industry are bubble-point determinations, pressure hold testing, and forward (diffusive air) flow measurements. These methods have the advantage of serving as in-process checks on the integrity of the sterilizing membrane while ensuring proper pore size. In the preparation of parenterals current practice is to subject the filter assembly to a nondestructive integrity test both prior to and after completion of the filtration operation.

The industry-accepted integrity tests are performed by applying gas, eg, air or nitrogen, pressure to a wetted filter and monitoring the air flow. These tests, applicable to both hydrophilic and hydrophobic membrane filters, include forward flow, bubble point, and pressure hold. The forward flow test quantitatively measures the diffusive flow as well as flow through any open pores in a wetted membrane filter. The test is performed by wetting a membrane filter and applying a predetermined constant pressure. The test pressure is established for a particular membrane filter by the filter manufacturer. The diffusional gas (air or nitrogen) flow rate as well as the flow through any open pores is measured through the wetted membrane. The gas flow is usually defined in units of mL min. The filter is integral if the forward flow value is lower than the manufacturer's specified value. The bubble point is qualitative and is dependent on the observation downstream of bubbling through the largest pores of a wetted filter. This test is usually employed for small surface area filters such as membrane disks. To perform the visual bubble point test, the filter membrane is first wet using an appropriate solvent, then air or nitrogen pressure is slowly applied until the wetting fluid is expelled from the largest pores and gas bubbles appear from a submerged tube in a downstream collection vessel. The pressure hold is a modified form of upstream forward flow testing and involves the measurement of decay in pressure over a specified time period for a given filter assembly and wetting fluid. Because the test is performed upstream, the downstream sterile connections are not disturbed. The advantage of this test is that it can be performed after sterilization of the filter assembly, as well as pre- and post-filtration. The filter housing is pressurized to the test pressure specified by the filter manufacturer, then the filter is isolated from the pressure source. The diffusion of gas across the wetted membrane is measured as a decay in pressure over a specified period of time. The pressure hold and forward flow tests are related through the ideal gas law (10). Automated filter integrity test instruments are available in order to provide reproducibility.

Validation Considerations. The need to validate all processes related to the manufacturing of pharmaceutical and biopharmaceuticals has been well established in good manufacturing practices (GMP) regulations and various other guidelines. Filtrative particle removal may be attributable to other mechanisms in addition to direct interception or size exclusion, including, for example, adsorptive particle arrest. It is therefore necessary to validate filter performance as the efficiency of the given filter is dependent on the physical, eg, viscosity and temperature, as well as the chemical, eg, presence of surfactants, composition of the suspending fluid. Microbial retention is required to be demonstrated under simulated pharmaceutical conditions in order to document the performance claims of the filter (3).

Validation of sterile filtration processes is required to be carried out under worst-case conditions. Typically, multiple filter lots (usually three) are challenged with the product under actual or simulated process conditions. A membrane filter having a pore size that allows penetration by the challenge organism is also included as a control. Most commonly, an 0.45 µm-rated filter is included when validation of 0.2 µm-rated sterilizing-grade filters are tested, and should show incomplete retention of *B. diminuta* (1). Essentially, the criteria for the selection of the challenge organism is that it should be small enough to challenge the retentivity of the sterilizing-grade filter and simulate the smallest organism that may occur during production.

If a specific organism has been identified as a routine bioburden, this organism may be substituted for *Brevundimonas (Pseudomonas) diminuta* (ATCC 19146). Ideally, validation experiments are conducted in the product under conditions that closely simulate process conditions. This is, however, not always possible. For example, prior to any microbial retention study it is necessary to determine viability of the test organism in the test fluid. If the product affects the viability of the test organism, as, for example, in the case of cytotoxic drugs, an appropriate substitution is essential. A placebo, ie, a formulation designed to simulate the actual product having the ionic strength, osmolality, viscosity, surface tension, and other attributes equivalent to the product, but which does not contain the active drug substance should be used. Alternatively, the native formulation may be modified, ie, the active ingredient is present but the bacteriocidal component removed neutralized diluted to ensure no effect on microbial viability. Similarly, if other process conditions affect the viability of the test organism, appropriate modifications are essential. Simulation of other process conditions is also essential. For example, hydraulic process conditions should be simulated during the bacterial challenge to assess any effect on the filter relative to its ability to retain bacteria. Such conditions include maximum differential pressure and pulsing. The pressure differential across the test filter should meet or exceed the maximum pressure differential observed during processing, within the design specifications provided by the filter manufacturer. This serves to validate the filter's ability to retain bacteria in the product and provide a sterile effluent up to or beyond the maximum process pressure differentials.

Sterilization Considerations. In sterile filtration processes the downstream side of the filter must be sterilized and must remain sterile during the entire process. Presterilized (gamma-irradiated) filters may be available or alternatively, filters may be sterilized by the user. The most common method of sterilization is by steam under pressure. The sterilization process must be validated to ensure that sterile conditions are met for a given system. Methods that involve steam are validated through the use of thermocouples and or biological indicators to ensure sterilization of the system. The filters may be sterilized in an autoclave. Alternatively, the sterilization may be undertaken in-place, called sterilization in-place (SIP) or *in situ* sterilization. Most filter manufacturers provide protocols and recommendations for these procedures. Small-volume systems tend to be autoclaved; larger systems are typically SIP. Minimally, a temperature of 121°C is used for autoclave sterilization; more commonly, a sterilization temperature of 125°C is employed. Temperatures in excess of 125°C may affect the plastics used for filter construction, thereby affecting the physical integrity of the membrane. Standard precautions must be followed during autoclaving: the system must be purged of air to achieve reliable sterilization, and the filter assembly wrapped using a porous barrier to ensure steam penetration. It is critical that excessive differential pressures are not created during the autoclaving which would result in damage to the membrane. These same considerations are all the more relevant during *in situ* sterilization. Additionally, in SIP operations the condensate must be drained throughout the steam cycle to prevent accumulation. This is achieved by keeping the drains and steam traps partially open during the steaming cycle. After the steam valve is closed, a noncondensable gas, such as air or nitrogen, is introduced into the housing to prevent formation of a vacuum on the upstream side of the filter. If it is necessary that the system be completely dry prior to use, the air–nitrogen flow can be continued until the system is dry and cool to operating temperature.

Other methods of sterilization may also be used. Irradiation has the advantage of assuring sterility without any residual gas components. However, several polymers used in filter manufacture may have limited resistance to irradiation sterilization. As with any other process, it has to be validated for sterility. Spores of *Bacillus pumilus* are the indicator organisms for validation of radiation sterilization. Another method is gas sterilization using a gas such as ethylene oxide. For successful sterilization of filters by ethylene oxide, the filters must be dry and wrapped so as to allow penetration of the gas into the matrix or the filter. In addition to the health and safety concerns associated with ethylene oxide gas itself, by-products such as ethylene chlorohydrin and ethylene glycol may be generated which also constitute health hazards. If ethylene oxide is used, appropriate venting is necessary following the sterilization cycle; however, in spite of venting there are concerns that some of the by-products may remain in the filter matrix. The sterilization method must be validated. The indicator organism recommended is *Bacillus subtilis* spores.

Sterile Filtration of Gases. Primary applications for sterile gas filtration are the sterilization of fermentor inlet air, fermentor vent gas, vents on water for injection tanks, and vacuum break filters during lyophilization. Operational and process considerations apply. Typically, the membrane in gas

filtration applications is a hydrophobic membrane, eg, PVDF or PTFE, although there are applications in which the liquid (condensate) in the system is well controlled and hydrophilic membranes may be used. The inherent hydrophobicity of membrane filters used for fermentor air sterilization allows these filters to remove bacteria completely from inlet air, even when exposed to moisture (11). The effluent for gas filtration applications is typically filtered at the 0.2- μm level.

Verification of the microbial retention efficiency of the membrane filters may be undertaken using either liquid or aerosol challenge tests. A liquid challenge test is more stringent. Furthermore, this test can provide retention information for process conditions such as extreme moisture after sterilization or air entrained with water drops. A liquid challenge is performed using a protocol similar to that described for liquid filtration.

Aerosol challenges may be conducted using essentially a test setup composed of three components: the nebulizer, mixing chamber, and a sampling system. An applicable system was first described in 1978 (12). A schematic representation of an aerosol challenge setup is provided in Figure 2. The challenge microorganism is aerosolized using a nebulizer. The aerosol is then mixed with compressed dry air to ensure that the monodispersed microbial challenge to the filter is delivered as a dry aerosol, rather than as microdroplets. Sampling may be done using a vacuum switch device that alternates between the upstream and downstream impingers, and a split-stream liquid impingement method. Any excess air flow not collected by the impingers is vented through an exhaust filter located upstream of the filter. Following the challenge, the buffer from the impingers located upstream and downstream of the test filters are assayed using standard microbiological methods. The dual impingers allow for precise determination of the actual challenge level for each test filter and calculation of the efficiency of titer reduction of the input challenge level.

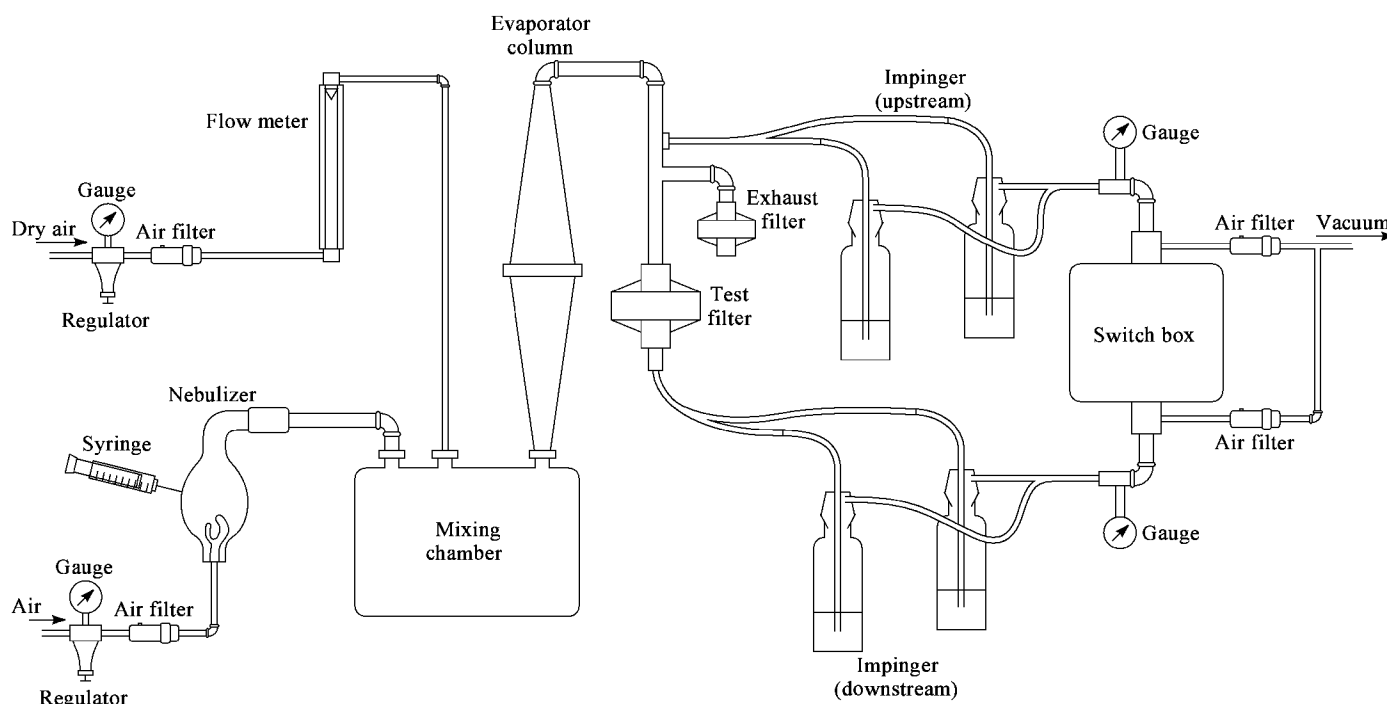


Fig. 2. Schematic of an experimental aerosol bacterial challenge setup for evaluation of bacterial retention.

Filters for use in sterile gas filtration must conform to standards similar to those mandated for sterile liquid filtration. Nondestructive integrity tests may be applied. The tests are performed by wetting the filter with an appropriate solvent, commonly 60–40 isopropyl alcohol–water for hydrophobic membranes, and applying air or nitrogen gas at a preset pressure.

Filtration of Virus Removal

General Principles. Filtration was traditionally used for the removal of bacteria and mycoplasma from biologicals that were heat labile. Advances in filtration technology have resulted in the availability of filtration devices for applications involving removal of viruses. The virological safety of biologicals and biopharmaceuticals is a key consideration in their manufacture. Much of the concern regarding viral contaminants in therapeutic agents centers around blood and blood products as well as biopharmaceuticals which have a blood or tissue component to their production. Human viruses of greatest concern have included human immunodeficiency virus (HIV), hepatitis B virus, hepatitis C virus, cytomegalovirus, and parvovirus. Nonhuman viruses such as bovine viral diarrhea virus are of concern if raw materials derived from these animals are incorporated into the production scheme. Viruses represent a diverse group which include enveloped and nonenveloped viruses, and ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) viruses of various sizes. The focus herein is on removal of viruses from fluids.

Methods used to ensure virological safety are briefly classified as either virus inactivation or virus removal methods. The former includes chemical inactivation, pasteurization, uv inactivation, and solvent–detergent and ion-exchange (qv) chromatography (qv) (13–16). Whereas these methods can be very effective depending on the inactivation process and the nature of the product, there are limitations to the application of these methods. Stabilizers, sometimes used to ensure that the biological activity is not compromised during the treatment, need to be removed from the final product. Heat treatment can denature certain proteins (17,18). Processes such as solvent–detergent, chemical treatment, and uv irradiation are often not uniformly effective against all viruses, especially enveloped ones.

The most desirable mechanism for the removal of viral particles using filtration is size exclusion. However, as in the case of bacterial removal by filtration, other mechanisms may also influence virus removal. These factors can include viral adsorption to the filter surface by electrostatic interactions; changes in pore size characteristics during filtration owing to deposits of material on the membrane surface, eg, development of a gel layer; and the filtration conditions, flow rate, pressure, temperature, etc. For example, adsorption can be a complex relationship between fluid pH, membrane chemistry, and the level of organics or protein in the fluid. Thus the removal of viruses by particle size minimizes many of the variables affecting the level of retention and can be a predictable means of sterilization. Removal of viral particles from fluids by size exclusion is preferable. However, concomitant with the requirement for adequate virus removal there is always the inherent necessity for no significant loss in product concentration and/or activity following filtration processes.

Virological Safety Considerations. Unlike bacterial removal where the specifications for a sterilizing-grade filter are clearly defined and have been implemented and validated over time, the virological safety of biologicals requires consideration of several complex issues. There are several inherent limitations of viral assays. Viral assays may lack the sensitivity to detect low levels of virus, which, although low, may be of medical concern. Demonstration of quantitative removal of bacterial contaminants can be made as, for example, by assay of the entire effluent post-filtration. It is not feasible to assay the entire volume for viral assays; therefore, proving absolute freedom from endogenous or adventitious viruses is impossible. Additionally, because of the specificity of virus assays, a different assay must be performed for each virus. Direct testing for the absence of viral contamination from a finished product is not considered sensitive enough for establishing freedom from an infectious virus. Historically, there have been several unfortunate iatrogenic accidents involving virus dissemination via administration of vaccines and blood products. Transmission of Hepatitis A (19), Hepatitis B (20), Hepatitis C (21), and

HIV (22) via administration of blood-derived products has been reported.

Minimization of risk of inadvertent exposure to real and theoretical viral contaminants is achieved by incorporation of multiple barriers to virus transmission in an integrated manner. This provides overlapping and complementary levels of protection with different modes of action, and separately derived virus reduction potential (23). Approaches that have been recommended are prevention of access of virus by screening of raw materials precursors used; monitoring production, ie, adventitious virus testing; and a general evaluation of the manufacturing process, ie, process validation of viral clearance. Establishing the absence of infectious virus in the final product often is not derived solely from direct testing for its presence but also from a demonstration that the purification regimen is capable of removal and or inactivation of viruses (24). Similar recommendations have also been made in Europe by the Committee for Proprietary Medicinal products (CPMP) (25).

Safety is thus the result of multiple barriers operating in concert. Whereas each approach individually may have limitations, use in an integrated manner provides overlapping and complementary levels of protection. These approaches may provide an effective method of overcoming risk and represent the only feasible approach in the face of theoretical risks which cannot be adequately characterized by classical technology.

Configurations of Virus Filtration Systems. The two principal membrane filtration systems for the removal of viral particles from fluids are single-pass or direct flow filtration and cross- or tangential flow filtration. In the first, the entire volume is filtered through the membrane filter. Typically, the membrane is either a flat sheet cut into disks or a pleated sheet assembled as a cartridge filter. The pore size of these types of membrane filters are generally between 50–100 nm (26). Pleating of the membrane has been used to increase filter surface area and thus increased volume and flow rate of the filtrate. Typically, filters in this format are designed to be used once.

The advantage of single-pass over cross-flow filtration is that it is an easier system to operate and can be cost effective, particularly if the product to be filtered is expensive, because very little of the initial fluid is lost during filtration. However, because the flow pattern of the fluid is directly through the filter, filter life may be too short for the fluid being filtered. The minimum flow rate needed downstream of the filter must also be considered, especially when there are time constraints to the process. In some situations it may be more advantageous to use a cross-flow system where higher flow rates may be easier to obtain.

In the cross-flow mode, fluid is passed across the membrane surface while a portion of the flow is diverted through the filter (permeate). A portion of flow is returned to the central reservoir as retentate. In this process the volume of fluid in the retentate continually decreases as more of the initial volume is collected as permeate. Viral particles are concentrated in the retentate. The advantage of this process is that the cross-flow across the membrane helps extend filter life by reducing gel layer formation. Typically the filtration systems utilizing cross-flow are either in the tangential-type system where fluid passes between two flat sheets of membrane material (Fig. 3) or consist of hollow-fiber filters where the fluid passes through the middle of hollow tubes (Fig. 4) (see HOLLOW FIBER MEMBRANES). For tangential flow, the membrane pore sizes range from 70–180 KD (27–29), as well as 100 and 300 KD polyethersulfone membranes. Hollow-fiber ultrafilters between 100 and 6 KD have been used for virus removal (30–33). These have been constructed from regenerated cellulose fiber, poly acetalnitrile (PAN), and polysulfone (PS). Both types of filters (tangential flow and hollow fiber) are typically designed for reuse after cleaning and sanitization and thus can be cost effective in terms of the filter cost. On the other hand, these systems may have higher holdup volumes and thus greater loss of production than single-pass systems. The cross-flow system may also be more complex and costly to install. Additionally, it is necessary to validate the filtration process. In the case of reuse of the filter system the cleaning and sanitization procedures also require validation.

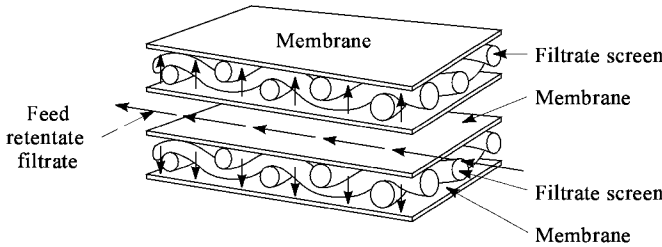


Fig. 3. Diagram of a tangential flow filtration system.

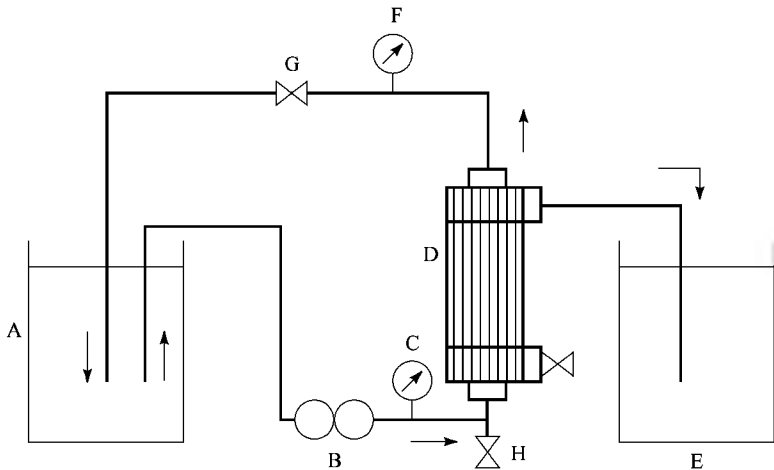


Fig. 4. Diagram of a hollow-fiber ultrafilter filtration system where A corresponds to the retentate reservoir; B, circulation pump; C, pressure gauge at module inlet; D, ultrafilter module; E, permeate reservoir; F, pressure gauge at module outlet; G, valve to control module outlet pressure; and H, drain valve.

Methods to Detect and Quantitate Viral Agents in Fluids. In order to assess the effectiveness of membrane filtration the ability to quantitate the amount of virus present pre- and post-filtration is critical. There are a number of techniques used. The method of choice for filter challenge studies is the plaque assay which utilizes the formation of plaques, localized areas in the cell monolayer where cell death caused by viral infection in the cell has occurred on the cell monolayer. Each plaque represents the presence of a single infectious virus. Virus quantity in a sample can be determined by serial dilution until the number of plaques can be accurately counted. The effectiveness of viral removal may be determined, as in the case of bacterial removal, by comparing the virus concentration in the input suspension to the concentration of virus in the effluent.

The plaque assay is desirable because it is very sensitive and only detects infectious viral particles. However, there are viral agents which cannot be supported by cell lines. In these cases other methods must be used. The polymerase chain reaction (PCR), which amplifies DNA or RNA from viral agents, can be used to detect the presence and quantity of viral agents. The amount of RNA or DNA target in the initial sample can be determined by competitive PCR where the quantity of amplified product is compared to a control PCR product where the initial amount of target is known. Quantification is also possible by an end-point dilution method similar to that used to determine a tissue culture infections dose. PCR methods can be very sensitive; however,

detection is based on the presence of nucleic acid and consequently the method does not differentiate between an infectious or noninfectious viral particle. In general, the detection of viral agents by either cell culture or PCR assay tends to be virus specific. Therefore multiple assay systems may be needed to detect more than one type of virus.

Effectiveness of Membrane Filtration.

Microfiltration. Various membrane filters have been used to remove viral agents from fluids. In some cases, membranes which have pores larger than the viral particle can be used if the filtration is conducted under conditions which allow for the adsorption of the viral particle to the membrane matrix. These are typically single-pass systems having pore sizes of 0.10–0.22 μm. Under situations which allow optimum adsorption, between 10–10² particles of poliovirus (28–30 nm) were removed (34–36). The formation of a cake layer enhanced removal (35). The titer reduction when using 0.10–0.22 μm membrane filters declined under conditions which minimized adsorption. By removal standards, these filters remove viruses at a rate on the low end of the desired titer reduction and the removal efficiency varies with differences in fluid chemistry and surface chemistry of viral agents (26).

Smaller pore size microfilters in single-pass systems which have pore sizes small enough to remove some viruses by size exclusion have been examined (26,37,38). Minimum levels of virus removal can be established for these systems if fluid and process conditions are employed which minimize removal of viral particles by mechanisms other than size selection.

Nylon filter membranes have been tested using 47-mm disks of the filter material (26,37). Influenza A virus (80–120 nm) and HIV (80–100 nm) were removed to below detectable limits in all fluids tested. However, removal of viruses smaller than influenza A virus was not as efficient. Titer reduction results for small (25–50 nm) viruses produced large differences depending on fluid type. The highest titer reductions were observed from high purity water and solutions having low concentrations of protein; smaller titer reductions were observed from solutions containing bovine serum. These results suggest that other factors in addition to size exclusion were enhancing the titer reduction for viruses in the 25–50 nm size range. Diminished adsorptive effects have been observed in the presence of serum or by pretreating a normally adsorbent membrane with serum or gelatin.

A PVDF membrane filter has been shown to remove >10⁶ particles of virus for viruses >50 nm independent of fluid type (8). Viruses smaller than 50 nm are not removed as efficiently but are removed in a predictable manner which correlates to the virus particle size. The chemistry of the suspending fluid affects titer reduction for viruses <50 nm owing to other removal mechanisms, such as adsorption, coming into play. The effects of these other mechanisms can be minimized by using filtration conditions that minimize adsorption.

Ultrafiltration. Ultrafilters have also been examined for viral removal by size exclusion utilizing tangential flow and hollow-fiber membrane systems (27–29,30–32,35,36). The titer reduction varies depending on virus size and membrane filter pore size distribution. Removal of poliovirus by a 30,000 molecular weight polysulfone ultrafilter removes >10⁴ particles of poliovirus in water of various qualities. Tangential flow ultrafiltration has been extensively tested using viruses ranging from 28 nm (Phi X174) to 80–100 nm (Murine leukemia virus) in size using a 70 KD PVDF membrane filter (28) (Table 1). There appeared to be a correlation between virus size and titer reduction. In this study, virus removal appeared to be enhanced in the presence of human serum albumin. Regenerated cellulose fiber (BMM) hollow-fiber ultrafilters from 10–80 nm have been tested using human blood borne viruses (30–32) (Table 1). Polysulfone (PS) and polyacrylonitrile (PAN) hollow-fiber ultrafilters have also been tested for their ability to remove viruses between 100–25 nm.

Table 1. Removal of Viral Particles from Fluids by Ultrafiltration

Filter		Virus (size)	Fluid ^a	Log titer reduction	Reference
Size reference	Material				
500K	PS ^b	MS2 (25 nm)	DI	~1.50	35
300K	ceramic ^c	MS2	DI	4.00	
100K	cellulosic ^d	MS2	DI	>6.00	
30K ~70 nm	PS ^e	polio (28–30 nm)	water	>6.00	28,29
		Phi X174 (28 nm)	water	>4.00	
	PVDF ^f	Phi X174	PBS	2.93	
		polio (28–30 nm)	PBS HSA	7.42	
		polio	PBS	3.1–3.51	
		SV-40 (40–45 nm)	PBS HSA	4.2	
		SV-40	PBS	4.89–5.65	
		Sindbis (54 nm)	PBS HSA	>5.7	
		Reo-3 (78 nm)	PBS	7.41	
		Reo-3	PBS	7.18	
		MuLV (85 nm)	PBS HSA	>7.6	
		HCV (35 nm)	PBS FCS	>6.82	
		HCV	plasma	1.50	32
105 nm	cellulose ^g	HCV	plasma	3.50	
		HCV	plasma	4.00	
		HIV (100 nm)	plasma	>4.00	30
		HIV	plasma	>5.77	
		HIV	plasma	>5.67	
		HIV	plasma	>5.85	
		HIV	plasma	>5.74	
		HIV	plasma	>5.94	
50,000	PAN ^h	polio (28–30 nm)	DMEM-10	4.59	33,36
		phage T1 ⁱ	DMEM-10	6.27	
13,000	PAN ^h	polio (28–30 nm)	DMEM-10	>6.51	
		phage T1 ⁱ	DMEM-10	7.55	
6,000	PAN ^h	phage PP7 (25 nm)	DMEM-10	7.62	
		phage T1 ⁱ	DMEM-10	7.80	
6,000	PS ^b	polio (28–30 nm)	DMEM-10	>6.40	
		phage T1 ⁱ	DMEM-10	7.99	

^a DI = deionized water; PBS, phosphate-buffered saline; HSA, human serum albumin; FCS, fetal calf serum; and DMEM, Dulbecco's Eagles minimum essential medium + 10% fetal bovine serum.

^b Polysulfone hollow fiber.

^c Tubular.

^d Hollow fiber.

^e Polysulfone flat sheet.
^f Poly(vinylidene fluoride), tangential flow.
^g Regenerated cellulose hollow fiber.
^h Polyacrylonitrile hollow fiber.
ⁱ 50-nm head, 150-nm tail.

Integrity Testing. As in the case of bacterial removal, it is necessary to carry out an integrity test on the filter, minimally, following filtration to ensure filter performance. Ideally, the integrity test should be performed both pre- and post-use. This is possible when a nondestructive integrity test method is used. Integrity tests used for virus removal filters include a forward-flow test similar to the test done on bacterial removal filters. This test is nondestructive and amenable to use both pre- and post-filtration. The test is correlated to virus removal by the filter manufacturer. Another nondestructive test is the liquid porosimetric integrity test, also correlated to virus removal. Another integrity test includes a gold particle removability test (GPT). This test is destructive and therefore is applicable only post-use. In general, the integrity test results must correlate with the virus removal claims, as specified by the filter manufacturer.

Validation Considerations. Mechanisms other than size exclusion may be operative in the removal of viruses from biological fluids. Thus virus removal must be validated within the parameters set forth for the production process and using membrane material representative of the product line of the filter.

The validation study for filtrative virus removal essentially involves challenging (spiking) the product using high titers of infectious virus under conditions that simulate process parameters and quantitating virus in pre- and post-treatment samples. Pre-purification treatments and post-purification modification reactions must also be validated. The choice of virus for validation studies is not as clearly defined as in bacterial filtration where there is an industry-accepted standard. No one viral agent can serve as a generic model virus. Thus, validation studies should be conducted using a panel of viruses that includes known contaminants which may represent identifiable and theoretical risks to product contamination, for example, HIV in the case of blood products; suspected contaminants or model viruses resembling suspected contaminants; or a range of viruses of differing properties which are not themselves considered likely contaminants. In some cases, use of surrogate viruses is necessary as some pathogenic viruses, eg, Hepatitis B and Hepatitis C, are not easily propagated in cell culture. Factors affecting virus clearance results include the choice of model viruses, the appropriateness of the scaled-down version, and the search for process variables which may alter the efficacy of virus inactivation–elimination steps, etc.

The virus reduction factor of an individual purification or removal–inactivation step is defined as the log₁₀ of the ratio of the virus load in the pre-purification material divided by the virus load in the post-purification material. A clearance factor for each stage can be calculated and the overall clearance capacity of the production process assessed. Total virus reduction is calculated as the sum of individual log reduction factors. Individual manufacturing steps must possess fundamentally different mechanisms of virus removal or inactivation in order for values to be considered cumulative. Additionally, because viruses vary greatly with regard to inactivation or removal profiles, only data for the same model virus can be cumulative.

A membrane filter which can uniformly remove all viral agents regardless of the size of the viral agent is not available. Part of the difficulty is that the efficient recovery of the biological product diminishes as the size difference between the virus and biological product lessens. Thus a balance needs to be met where virus removal and product recovery are optimized.

Improvements in membrane technology, validation of membrane integrity, and methods to extend filter usage should further improve the performance of membrane filters in removal of viral particles. Methods to improve or extend filter life and increase flow rates by creating more complex flow patterns could possibly be the focus of the next generation of membrane filters designed to remove viral particles.

BIBLIOGRAPHY

1. F. W. Bowman, M. P. Calhoun, and M. White, *J. Pharm. Sci.* **56**, 222 (1967).
2. M. Osumi, N. Yamada, and M. Toya, *J. Pharm. Sci. Technol.* **50**, 30 (1996).
3. *Guidelines on Sterile Drug Products Produced by Aseptic Processing*, Center for Drugs and Biologics and Office of Regulatory Affairs, U.S. FDA, Washington, D.C., June 1987.
4. "Indirect Food Additives Subpart B: Substances for Use as Basic Components of Single and Repeat Use Food Contact Surfaces," *Code of Federal Regulations*, Title 21, Part 177, U.S. Government Printing Office, Washington, D.C., 1994.
5. *Standard Test Method for Determining Bacterial Retention of Membrane Filters Utilized for Liquid Filtration*, ASTM F838-83, American Society for Testing and Materials, Philadelphia, Pa., 1983; reapproved 1988.
6. "Current Good Manufacturing Practice in Manufacturing Processing, Packing or Holding of Drugs: General," *Code of Federal Regulations*, Title 21, Part 210.3 (b) (6), U.S. Government Printing Office, Washington, D.C., 1994.
7. "Current Good Manufacturing Practice for Finished Pharmaceuticals," *Code of Federal Regulations*, Title 21, Part 211.72, U.S. Government Printing Office, Washington, D.C., 1994.
8. H. Aranha and J. Meeker, *PDA J. Pharm. Sci. Technol.* **42**, 67 (1995).
9. J. Meeker, J. B. James, J. M. Martin, and G. Howard, Jr., Technical data, STR RUF08BA, Pall Corp., East Hills, N.Y., 1990.
10. *Validation Guide for the Palltronic™ FFE03-P Filter Integrity Test Instrument*, TRFFE03-P, Pall Corp., East Hills, N.Y., 1992.
11. C. F. Bruno and L. A. Szabo, *Biotechnol. Bioeng.* **25**, 1223 (1983).
12. R. Duberstein and G. Howard, *J. Paren. Drug. Assoc.* **32**, 192 (1978).
13. E. D. Gomperts, *Lancet* **23**, 295 (1986).
14. M. S. Horowitz, C. Rooks, B. Horowitz, and M. Hilgortner, *Lancet ii*, **186** (1988).
15. P. M. Mannucci and M. Colombo, *Lancet*, 782 (1988).
16. P. Murphy, T. Nowak, S. M. Lemon, and J. Hilfenhaus, *J. Med. Virol.* **41**, 61 (1993).
17. M. Gleeson, L. Herd, and C. Burns, *Ann. Clin. Biochem.* **27**, 592 (1990).
18. A. M. Prince, B. Horowitz, M. S. Horowitz, and E. Zang, *Eur. J. Epidemiol.* **3**, 103 (1987).
19. Centers for Disease Control, *Morbid. Mortal. Weekly Rep.* **45**, 29 (1996).
20. R. D. Aach and R. A. Kahn, *Ann. Intern. Med.* **92**, 539–546 (1980).
21. Y. J. Wang and co-workers, *Vox Sang.* **67**, 187–190 (1994).
22. S. C. Darby, D. W. Ewart, P. L. Giangrande, P. J. Dolin, R. J. D. Spooner, and C. R. Rizza, *Nature*, **377**, 79 (1995).
23. S. Builder, R. Van Reis, N. Paoni, and J. Ogez, in R. W. Spier, J. B. Griffiths, J. Stephenne, and P. J. Crooy, eds., *Advances in Animal Cell Biology and Technology for Bioprocesses*, Butterworths, Stoneham, Mass., 1989, p. 452.
24. "Viral Safety Evaluation of Biotechnology Products Derived from Cell Lines of Human or Animal Origin," ICH Harmonized Tripartite Guideline, IFPMA, Geneva, Switzerland, 1997, pp. 1–27.
25. "Note for Guidance on Plasma Derived Medicinal Products," CPMP BWP 269 95, Committee for Proprietary Medicinal Products (CPMP), London, Mar. 13, 1996.
26. K. H. Oshima, A. K. Highsmith, and E. G. Ades, *Environ. Topical. Water Qual.* **9**, 165 (1994).
27. A. J. DiLeo, A. E. Allegranza, and S. E. Builder, *Bio/ Technol.* **10**, 182 (1992).
28. A. J. DiLeo, D. A. Vacante, and E. F. Deane, *Biologicals*, **21**, 275 (1993).

29. Ref. 28, p. 287.
30. Y. Hamamoto and co-workers, *Vox Sang*, **56**, 230 (1989).
31. S. Sekiguchi and co-workers, *Membrane*, **14**, 253 (1989).
32. T. Yuasa and co-workers, *J. Gen. Virol.* **72**, 2021 (1991).
33. K. H. Oshima and co-workers, *Can. J. Microbiol.* **41**, 316 (1995).
34. K. Hou, C. P. Gerba, S. M. Goyal, and K. S. Zerda, *Appl. Environ. Microbiol.* **40**, 892 (1980).
35. J. G. Jacangelo, S. S. Adham, and J-M. Laine, *J. AWWA*, **87**, 107 (1995).
36. S. S. Madaeni, A. G. Fane, and G. S. Grohman, *J. Membr. Sci.* **102**, 65 (1995).
37. K. H. Oshima and co-workers, *J. Acquir. Immune Defic. Syndr.* **8**, 64 (1995).
38. K. H. Oshima, T. T. Evans-Strickfaden, and A. K. Highsmith, *Biologicals* **124**, 137 (1996).

General References

H. Aranha-Creado, K. Oshima, S. Jafari, G. Howard, Jr., and H. Brandwein, *Scientific and Technical Report*, STR-PUF-24, Pall Ultrafine Filtration Co., East Hills, N.Y., 1995.
"Bacterial Endotoxins Test <85>," "Biological Reactivity Tests, *in vivo* <88>," "Particulate Matter in Injections <788>," and "Purified Water," *USP 23*, The U.S. Pharmacopeial Convention, Rockville, Md., 1994.
T. J. Leahy and M. J. Sullivan, *Pharm. Technol.* **2** (11).
R. V. Levy and T. J. Leahy, in S. S. Block, *Disinfection, Sterilization, and Preservation*, 4th ed., Lea & Febiger, Philadelphia, Pa., 1991, pp. 527–551.
P. Keating and co-workers, *BioPharm* **5** (1992).

Hazel Aranha-Creado
K. Oshima
Pall Corporation

MICROEMULSIONS

There is no official or universally accepted definition of what constitutes a "microemulsion." In fact, for several years, some leading scientists in microemulsion research considered the term to be an unnecessary and even an unfortunate one. Nevertheless (Table 1), during the years from about 1975 to 1980 the word ascended from obscurity to ubiquity. By the end of 1996 there were 13 widely available English-language books (1–9) with the word "Microemulsion" in their titles (10). About 70 more books on surfactants are in print, of which those on industrial applications (9,11–18), and environmental effects (19–21) are of particular interest here.

Table 1. Growth in Microemulsion Publications

Year	Number ^a of publications
1970	12
1975	29
1980	123
1985	237
1990	410

^a In *Chemical Abstracts*.

The concept of microemulsions now holds a central role within the field of surfactant technology. Perhaps the most fundamental fact captured by the term is that, contrary to a popular saying, oil and water can mix.

Definition of a Microemulsion

The term microemulsion was introduced by Schulman, who studied surfactant solutions as early as 1943 (22). At that time it was widely accepted that "oil and water do not mix," and Schulman understood that an emulsion scatters light because it contains droplets whose diameters are large compared to the wavelength of light (see EMULSIONS). Thus, the term *microemulsion* implies a system which (like an emulsion) contains droplets of oil or water, but in which the droplets are too small to scatter light.

Although the complete description of what constitutes a microemulsion was controversial for some years and some differences of usage still remain, virtually all researchers now agree on certain qualifiers. A microemulsion is a true, thermodynamically stable, liquid solution that contains water, oil, and at least one amphiphile. Many microemulsions also contain one or more inorganic salts. Typically the oil is a mineral oil or hydrocarbon, but it may be almost any nonpolar compound. The oil and the amphiphile may be single, pure components; or (as in most commercial formulations), the oil, amphiphile, or both may contain an indefinitely large number of compounds. Some researchers insist that a microemulsion must contain at least one true surfactant (eg, an amphiphile that forms liquid crystalline phases and has a well-defined critical micelle concentration). Many widely used microemulsions contain both a true surfactant and a cosurfactant (sometimes called a cosolvent), ie, an amphiphile whose molecular weight is too small for it to be a true surfactant. However, some members of a homologous series of amphiphiles may be surfactants and others not, depending only on their molecular weight. For example, depending on the values of *m* and *n*, the amphiphile C_{*m*} H_{*2m+2*} (OC₂H₄)_{*n*} OH may be an alcohol, a cosolvent, or a true surfactant. Moreover, many microemulsion properties change only incrementally when a small change is made in the molecular weight of the amphiphile. Therefore, some researchers consider that almost any amphiphile–oil–water solution may be fairly called a microemulsion.

Many solutions of common nonionic surfactants and water separate into two phases when heated above a certain temperature (the cloud point), and some investigators call the phase of greater surfactant concentration, a microemulsion. Thus, there is not even universal agreement that a microemulsion must contain oil.

Microemulsions and Phase Diagrams

From the very beginning (22), the existence or nonexistence of a microemulsion was seen to depend not only on the presence of certain classes of compounds (ie, components), but also on the concentrations of these components. The number of phases present and their compositions, when presented in graphical form, constitute a phase diagram. Moreover, as specified by Gibbs' phase rule, amphiphile–oil–water–phase diagrams form characteristic patterns that change in qualitatively similar ways when the temperature, pressure, concentrations, or molecular structures of the components are changed. Thus, phase diagrams offer not only another way to define microemulsions, but also a rigorous way to clarify differences in terminology and usage.

Figure 1 (23) illustrates the phase diagram of an amphiphile–oil–water system (24) such as C₄H₉OC₂H₄OH ("C4E1")–decane–water (25) or C H₁₂–(OC₂H₄)₂OH (C E₂)–tetradecane–water (26). For a real surfactant, such as C₁₂E₄, the diagram would be more complicated, because of the occurrence of liquid crystalline phases (27). Samples whose compositions fall within the tietriangle of Figure 1 form three liquid phases, of compositions, *T*, *M*, and *B* (corners of the tietriangle). Each pair of adjacent corners of the tietriangle is connected by a binodal curve (as well as by a side of the tietriangle). Compositions between a side of the tietriangle and the adjacent binodal curve form two conjugate phases in equilibrium with each other; each such pair of phases is connected by a tieline. For two of the binodals the compositions of the conjugate phases can become closer and closer and their connecting tielines shorter and shorter, until the phase compositions and the end points of the tielines become identical at a plait point. Any composition outside of the tietriangle and the three binodal curves forms only a single liquid phase. On the amphiphile–water side of the phase diagram these single-phases contain only amphiphile and water (no oil); on the amphiphile–oil side of the phase diagram these single-phases contain only amphiphile and oil (no water).

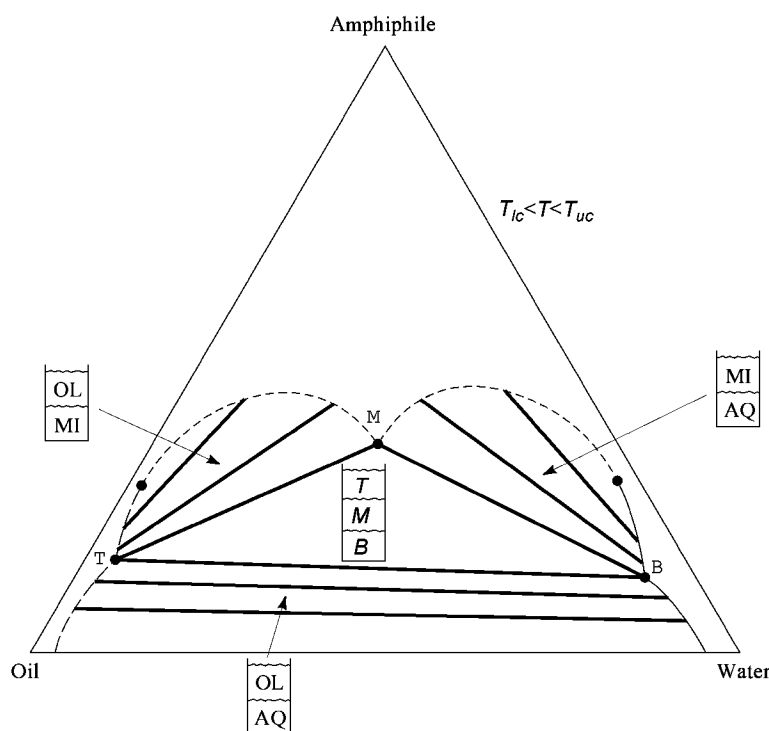


Fig. 1. Phase diagram of an amphiphile–oil–water system that forms a middle-phase microemulsion, definition of microemulsion, and illustration of the pairs (and triad) of phases formed in the various multiphase regions of the diagram. Boundaries: (—), aqueous (AQ); (---), oleic (OL); (---), limiting microemulsion (MI).

By the most general definition of a microemulsion, every phase described by Figure 1 would be a microemulsion. But, clearly, when two (or even three) phases are simultaneously present, little (except confusion) is gained by giving the different phases the same name. Accordingly, along the binodal curves (which describe the compositions of conjugate phases) only compositions between the two plait points are termed microemulsions. Other conjugate phases are called oleic phases or aqueous phases, respectively, depending on whether their main component is oil or water. (Oleic phases and aqueous phases are ubiquitously called oil and water, even in scientific journals. This practice should be rigorously avoided, because it produces unnecessary confusion between components, at the corners of the phase diagrams, and phases, which always contain at least minute concentrations of all components present in the system.)

In Figure 1, the pairs (or triad) of phases that form in the various multiphase regions of the diagram are illustrated by the corresponding test-tube samples. Except in rare cases, the densities of oleic phases are less than the densities of conjugate microemulsions and the densities of microemulsions are less than the densities of conjugate aqueous phases. Thus, for samples whose compositions lie within the oleic phase–microemulsion binodal, the upper phase (ie, layer) is an oleic phase and the lower layer is a microemulsion. For compositions within the aqueous phase–microemulsion binodal, the upper layer is a microemulsion and the lower layer is an aqueous phase. When a sample forms two layers, but the amphiphile concentration is too low for formation of a middle phase, neither layer is a microemulsion. Instead the upper layer is an oleic phase ("oil") and the lower layer is an aqueous phase ("water").

In three-phase systems the top phase, *T*, is an oleic phase, the middle phase, *M*, is a microemulsion, and the bottom phase, *B*, is an aqueous phase. Microemulsions that occur in equilibrium with one or two other phases are sometimes called "limiting microemulsions," because they occur at the limits of the single-phase region.

Rigorous and useful as it is, the definition of limiting microemulsions does not specify which compositions in the single-phase region should also be called microemulsions. In Figure 1 there is critical micelle concentration point for formation of normal micelles on the amphiphile–water side of the phase diagram and an inverse critical micelle concentration point for formation of inverted micelles on the amphiphile–oil side of the diagram (28). (Depending on the compounds and temperature, these CMC points may be easily detected, or exist only in a formal sense.) Geometrically, each of the two CMC points can be connected by a line to a nearby plait point. These lines serve to separate the single-phase area of the diagram formally into (nonlimiting) microemulsion, aqueous, and oleic regions. However, such formalities seldom are needed; and sometimes they may even be counterproductive.

Temperature and Salinity Scans

The locations of the tietriangle and binodal curves in the phase diagram depend on the molecular structures of the amphiphile and oil, on the concentration of cosurfactant and or electrolyte if either of these components is added, and on the temperature (and, especially for compressible oils such as propane or carbon dioxide, on the pressure (29,30)). Unfortunately for the laboratory worker, only by measuring (or correctly estimating) the compositions of *T*, *M*, and *B* can one be certain whether a certain pair of liquid layers are a microemulsion and conjugate aqueous phase, a microemulsion and oleic phase, or simply a pair of aqueous and oleic phases.

However, often the identities (aqueous, oleic, or microemulsion) of the layers can be deduced reliably by systematic changes of composition or temperature. Thus, without knowing the actual compositions for some amphiphile and oil of points *T*, *M*, and *B* in Figure 1, an experimentalist might prepare a series of samples of constant amphiphile concentration and different oil–water ratios, then find that these samples formed the series (a) 1 phase, (b) 2 phases, (c) 3 phases, (d) 2 phases, (e) 1 phase as the oil–water ratio increased. As illustrated by Figure 1, it is likely that this sequence of samples constituted (a) a "water-continuous" microemulsion (of normal micelles with solubilized oil), (b) an upper-phase microemulsion in equilibrium with an excess aqueous phase, (c) a middle-phase microemulsion with conjugate top and bottom phases, (d) a lower-phase microemulsion in equilibrium with excess oleic phase, and (e) an oil-continuous microemulsion (perhaps containing inverted micelles with water cores).

Historically, however, it has been much more common for experimentalists to introduce a new variable into Figure 1, changing either the temperature of one or more samples of fixed composition, or the electrolyte concentration in a series of samples of fixed amphiphile–oil–water ratio. The former constitutes a temperature scan; the latter experiment is widely known as a salinity scan. When the temperature of an amphiphile–oil–water system is varied, the phase diagram can be plotted as a triangular prism (because temperature is an intensive or field variable). When a fourth component (eg, NaCl) is added at constant temperature, tetrahedral coordinates, are appropriate (conjugate phases have different salinities, and the planes of different tietriangles are no longer parallel).

Changes of temperature or salinity cause the phase diagram to evolve through the sequence illustrated by Figure 2. As the temperature is decreased from some temperature between T_c and T_{uc} the aqueous phase–microemulsion binodal shrinks until it becomes a single point, the lower critical endpoint. Simultaneously, the tietriangle shrinks in height, becoming a lower critical tieline of zero height at $T = T_{lc}$. Conversely, when the temperature is raised, the

oleic phase-microemulsion binodal shrinks, until it becomes an upper critical end point at T_{uc} . Below T_{uc} and above T_{uc} the respective binodal curves and plait points disappear, leaving only a single two-phase region. Figure 2 illustrates the phase diagrams of a nonionic amphiphile–oil–water system at $T < T_{lc}$ and $T > T_{uc}$, respectively. Because of the relative densities of the phases, for $T < T_{lc}$ the system is sometimes said to contain an upper phase microemulsion in equilibrium with an aqueous phase, whereas for $T > T_{uc}$ the system may be said to contain a lower phase microemulsion in equilibrium with an oleic phase. However, if the existence of middle-phase microemulsions at intermediate temperatures is unknown, the respective phase pairs are likely to be called simply oil and water.

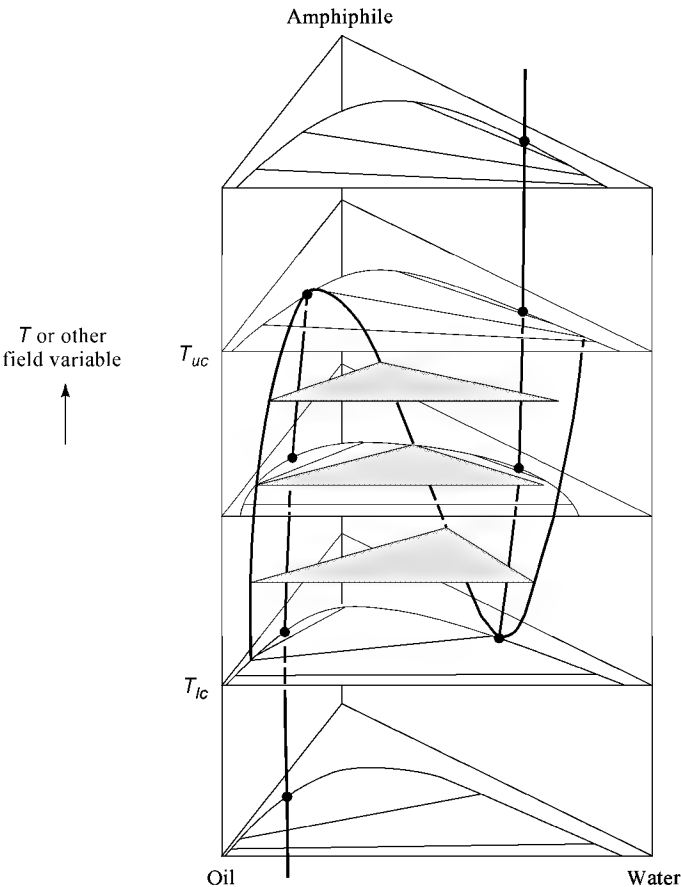


Fig. 2. The phase diagrams and terminology of a microemulsion system close to its two critical end points, where the middle phase and one of the binodals disappear.

Of all the characteristic points in the phase diagram, the composition of the middle phase is most sensitive to temperature. Point M moves in an arc between the composition of the bottom phase (point B) at T_{lc} and the composition of the top phase (point T) at T_{uc} , reaching its maximum surfactant concentration near $T = (T_{lc} + T_{uc}) / 2$. (Points B and T move by much smaller amounts, also.) The complete nonionic-amphiphile–oil–water–temperature phase diagram is illustrated by Figure 3, including the S-shaped curve of T , M , and B compositions; the two lines of plait points, which terminate at the lower and upper critical end points; and the lower and upper critical tielines (at T_{lc} and T_{uc} , respectively).

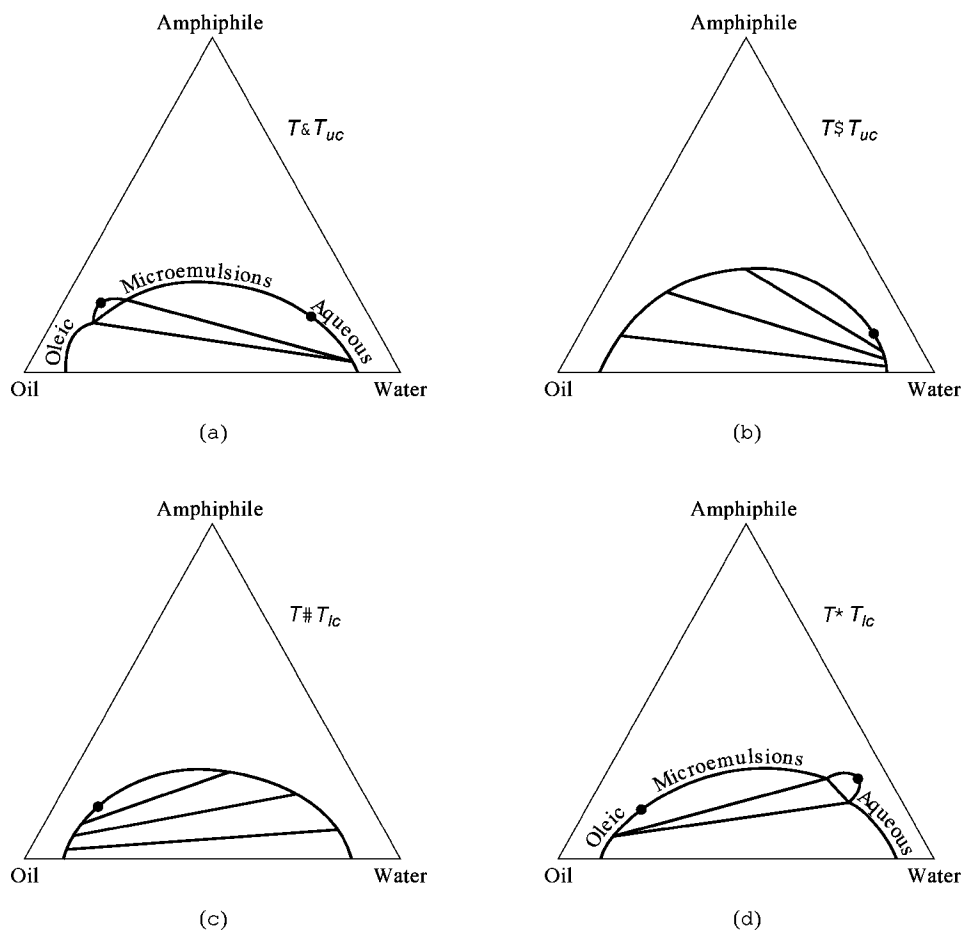


Fig. 3. Typical nonionic amphiphile–oil–water–temperature phase diagram, illustrating (a) the S-shaped curve of T , M , and B compositions, (b) the lines of plait points, (c) the lower and upper critical end points (at T_{lc} and T_{uc} , respectively), and (d) the lower and upper critical tielines.

Figure 4 illustrates the analogous amphiphile–oil–water–electrolyte phase diagram, including a representative tetrahedron, the S-shaped curve of T , M , and B compositions, the lower and upper critical end points (R and Q, respectively), and the lower (PR) and upper (QS) critical tielines (31). For clarity, the two-phase regions on each side of the stack of tetrahedrons have been omitted; the effects of salinity on the binodals are analogous to the temperature effects illustrated by Figure 2.

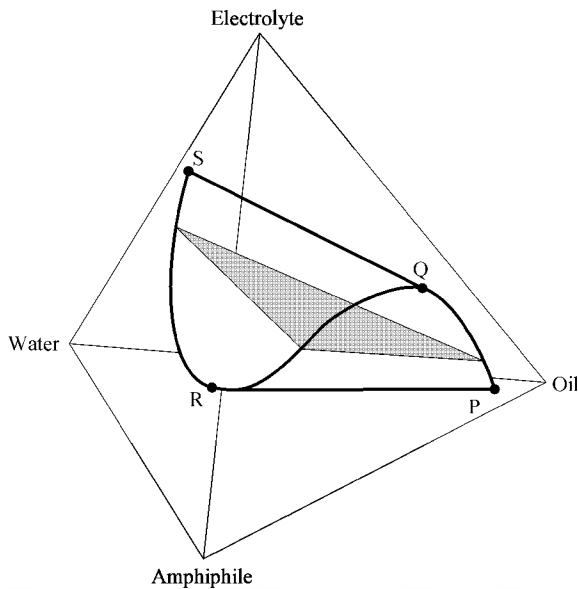


Fig. 4. Typical amphiphile–oil–water–electrolyte phase diagram, illustrating the S-shaped curve of T , M , and B compositions, the lower and upper critical end points (R and Q, respectively), and the lower (PR) and upper (QS) critical tielines (31).

For nonionic amphiphiles, the effects of temperature on the phase behavior are large and the effects of inorganic electrolytes are very small. However, for ionic surfactants temperature effects are usually small, but effects of inorganic electrolytes are large. Most common electrolytes (eg, NaCl) "salt out" ionic surfactants, making them less soluble in water and moving M from B to T . However, a few electrolytes "salt in" ionic surfactants, and have the opposite effect (32).

When the phase behavior of the amphiphile–oil–water–electrolyte system of Figure 4 is measured at different temperatures (31), the resulting phase diagram has four dimensions. Figure 5 (33) illustrates the effect of temperature on the positions of the critical tielines (which mark the limits of existence of middle-phase microemulsion) within the tetrahedral diagram. Also plotted in Figure 5 is the temperature vs the salt concentrations (of the midpoints) of the critical tielines. For this particular system, as the temperature of the phase diagram measurements is increased, the lower and upper salinity limits of existence of the middle phase both decrease, but the two critical tielines meet at a cusp (in this example, at a physically unrealizable negative salinity). This is called a tricritical point, because at it the tetrahedron of phases T , M , and B shrinks to a single point (34). Tricritical points are the thermodynamic origin of ultralow interfacial tensions, which are exploited in micellar-polymer enhanced oil recovery (EOR) and in surfactant-enhanced aquifer remediation (SEAR)

analogue.

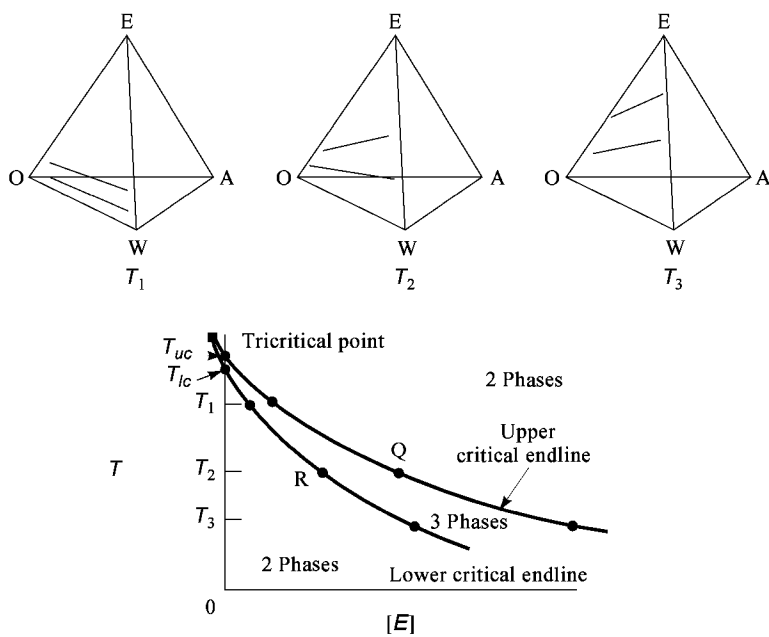


Fig. 5. Lower and upper critical tielines in a quaternary system at different temperatures; and a plot of the critical end point salinities vs temperature, illustrating lower critical endline, upper critical endline, optimal line, and tricritical point for four-dimensional amphiphile–oil–water–electrolyte-temperature phase diagram (33).

Microemulsions of ionic surfactants contain ionic surfactant, as well as nonionic surfactant or amphiphile, oil and water, and inorganic electrolyte. Their phase behavior and thermodynamics are similar to those described above for nonionic surfactant–oil–water–electrolyte systems, except that ionic-surfactant systems contain an additional degree of freedom due to the extra component. The effects of the organic salt (ie, ionic surfactant) are analogous to those of inorganic salts. Hence, diagrams similar to Figure 5 can be measured, in which inorganic electrolyte is replaced by organic electrolyte. The phase diagrams have five dimensions (four fractional concentrations and temperature), not just four; hence, instead of a single tricritical point, they have a line of tricritical points. The dimensionality of such systems makes them difficult to understand and study. But the extra degree of freedom can be used, when needed, as an additional variable to tune the system and its physical properties to meet the engineering requirements of an application (33).

Physical Properties and Applications

The current or potential industrial applications of microemulsions include metal working, catalysis, advanced ceramics processing, production of nanostructured materials (see NANOTECHNOLOGY), dyeing, agrochemicals, cosmetics, foods, pharmaceuticals, and biotechnology (9,12–18). Environmental and human-safety aspects of surfactants have begun to receive considerable attention (19–21).

Microemulsions became well known from about 1975 to 1980 because of their use in "micellar-polymer" enhanced oil recovery (EOR) (35). This technology exploits the ultralow interfacial tensions that exist among top, microemulsion, and bottom phases to remove large amounts of petroleum from porous rocks, that would be unrecoverable by conventional technologies (36,37). Since about 1990, interest in the use of this property of microemulsions has shifted to the recovery of chlorinated compounds and other industrial solvents from shallow aquifers. The latter application (15) is sometimes called surfactant-enhanced aquifer remediation (SEAR).

Figure 6 illustrates how the three tensions among the top, middle, and bottom phases depend on temperature for a system of nonionic surfactant–oil–water (38), or on salinity for a representative system of anionic surfactant–cosurfactant–oil–water and electrolyte (39). As T approaches T_{uc} from lower temperatures, the composition of M approaches the composition of T , and the interfacial tension between them, σ_{MT} , goes to zero at $T = T_{uc}$. Similarly, as T approaches T_{lc} from higher temperatures, the composition of M approaches the composition of B , and the interfacial tension between them, σ_{MB} , goes to zero at $T = T_{lc}$. If Antonov's rule is obeyed (it is virtually always an excellent approximation for applied work) (40,41), the tension between the nonmicroemulsion pair of phases is given by

$$\sigma_{BT} = \sigma_{MB} + \sigma_{MT}$$
 (1)

The temperature (or salinity) at which $\sigma_{MB} = \sigma_{MT}$ is called the optimal temperature (or optimal salinity), because at that temperature (or salinity) the oil–water interfacial tension is a minimum, which is optimum for oil recovery. For historical reasons, the optimal temperature is also known as the HLB (hydrophilic–lipophilic balance) temperature (42,43) or phase inversion temperature (PIT) (44). For most systems, all three tensions are very low for $T_{lc} < T < T_{uc}$, and the tensions of the middle-phase microemulsion with the other two phases can be in the range 10^{-5} – 10^{-7} N/m. These values are about three orders of magnitude smaller than the interfacial tensions produced by nonmicroemulsion surfactant solutions near the critical micelle concentration. Indeed, it is this huge reduction of interfacial tension which makes micellar-polymer EOR and its SEAR counterpart physically possible.

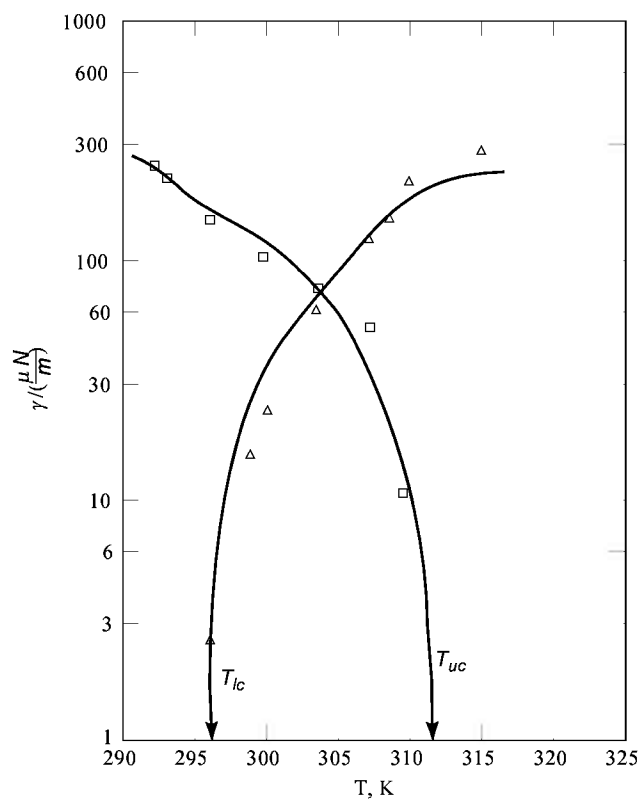


Fig. 6. Interfacial tensions among microemulsion, top, and bottom phases.

The shape of the σ_{MB} and σ_{MT} curves is theoretically well established by critical scaling theory. The temperature dependence is given by

$$\sigma_{MJ} = \sigma_{T,J} \left| \frac{T - T_C}{T_C} \right|^\gamma \tag{2a}$$

and at constant temperature the composition dependence is

$$\sigma_{MJ} = \sigma_{C,J} \left| \frac{C - C_C}{C_C} \right|^{2\nu(1-\alpha)} \tag{2b}$$

Here σ_{TJ} and σ_{CJ} are nonuniversal parameters, whose values must be measured for each chemical system. However, exponents such as α , β , γ , and ν are universal parameters, whose values are known from theory to be $\alpha = 0.89$, $\beta = 0.345$, $\gamma = 1.29$, and $\nu = 0.64$ for all systems (45). Here J is B or T , T_C is the appropriate critical end point temperature, in K, and C_C is the critical-point concentration.

Modern scaling theory is a quite powerful theoretical tool (applicable to liquid crystals, magnets, etc) that has been well established for several decades and has proven to be particularly useful for multiphase microemulsion systems (46). It describes not just interfacial tensions, but virtually any thermodynamic or physical property of a microemulsion system that is reasonably close to a critical point. For example, the compositions of a microemulsion and its conjugate phase are described by equations of the following form:

$$C_J - C_I = C_1 \left| \frac{\mu - \mu_C}{\mu_C} \right|^\beta \tag{3a}$$

$$C_J + C_I = C_2 \left| \frac{\mu - \mu_C}{\mu_C} \right|^{1-\alpha} \tag{3b}$$

where μ is the chemical potential and C_1 and C_2 depend on the combination of components. Equations 2 and 3 each contain one or more nonuniversal parameters, whose values depend on the compounds and must be measured, and universal exponents, whose values are the same for all systems and are known. These equations provide two principal practical advantages for fitting experimental data: (a) unlike empirical equations, the forms of the equations already are known and have a theoretical basis; (b) the values of the exponents in the equations are accurately known, and this reduces the number of fitting parameters.

The values of the exponents for ordinary critical points or bicritical points (where two phases become identical) are called nonclassical, because (unlike the exponents in van der Waals and other classical equations) they are not multiples of $1/2$.

The salinities of connate brines in oil reservoirs range from potable to saturated at elevated temperature and pressure. When a reservoir of high salinity has been flooded previously with fresh water, the brine salinity can vary greatly within the reservoir. To find a suitable surfactant requires a laborious search for a formulation with the correct optimal salinity for each reservoir. Thus (see Fig. 6), there has been a desire to find a surfactant that would form three phases over the broadest range of salinities and have ultralow interfacial tensions for those phases at the same time. However, there are thermodynamic limits on the extent to which these goals can be simultaneously met. These limits can be qualitatively understood by examination of Figure 3. If the two microemulsion interfacial tension lines are moved (horizontally, without change of shape) so that the critical end points are very close together, the tensions where the two curves intersect (ie, the optimal salinity) are all very low, but the width of the three-phase region (ie, distance between the critical points) is very small. If the two curves are slid sideways so the critical points are far apart, the range of three-phase salinities is very wide, but nowhere are all three interfacial tensions ultralow.

Changing the distance between the critical points requires a new variable (in addition to the three independent fractional concentrations of the four-component system). As illustrated by Figure 5, the addition of a fourth thermodynamic dimension makes it possible for the two critical end points to approach each other, until they occur at the same point. As the distance between the critical end points decreases and the height of the stack of tietriangles becomes smaller and smaller, the tietriangles also shrink. The distance between the critical end points (see Fig. 5) and the size of the tietriangles depend on the distance from the tricritical point. These dependencies also are described scaling theory equations, as are physical properties such as interfacial

tension.

However, as given by group renormalization theory (45), the values of the universal exponents depend on the (thermodynamic) dimensionality of the system. For four dimensions (as required by the phase rule for the existence of tricritical points), the exponents have classical values. This means the values are multiples of 1/2. The dimensions of the volume of tie-triangles are (31)

$$h-h_O\left|\frac{T-T_{tc}}{T_{tc}}\right|^{3/2}$$

(4a)

$$w-w_O\left|\frac{T-T_{tc}}{T_{tc}}\right|^{2/2}$$

(4b)

$$l-l_O\left|\frac{T-T_{tc}}{T_{tc}}\right|^{1/2}$$

(4c)

As confirmed by experiment (47), the dependence of the optimal tension on the distance from the tricritical point is given (40) by

$$\sigma-\sigma_O\left|\frac{T-T_{tc}}{T_{tc}}\right|^{4/2}$$

(5)

where T_{tc} is the (absolute) tricritical point temperature and the values of h_O , l_O , w_O , and σ_O depend on the chemical system. Thus, there is a simple, universal, thermodynamic relationship between the width of the three-phase region, the composition (ie, amount of solubilized oil or brine) of the middle-phase microemulsion at optimum, and the optimal tensions. Once the values of h_O , l_O , w_O , T_{tc} , and σ_O are known, equations 4 and 5 provide convenient equations for the description of the compositions of phases T , M , and B and the interfacial tensions among them.

Moreover, except for small tradeoffs between h_O and σ_O that may be obtained by changing the surfactant, the goals of simultaneously lowering the tensions and increasing the width of the three-phase region are mutually contradictory. As in the design of heat engines, thermodynamics can save some work in the design of microemulsion formulations (33)!

Microemulsions and Macroemulsions

Operationally, it is not always easy to determine whether a given sample is a microemulsion or macroemulsion. Close to a critical point, a microemulsion may not be transparent, but may be confused with a macroemulsion. An extremely stable macroemulsion may require more than one researcher's lifetime to separate into bulk layers, and thus, appear to be a microemulsion, but it is nevertheless a (macro) emulsion. Such experimental problems added to the early difficulties of developing the microemulsion concept. When the number of components is so large that determination of the phase diagram is impractical, advanced techniques such as quasi-elastic light scattering, small-angle x-ray or neutron diffraction, or nuclear magnetic resonance may be used to determine if the sample is a microemulsion by characterization of its structure (2).

However, the formal differences between microemulsions and macroemulsions are well defined. A microemulsion is a single, thermodynamically stable, equilibrium phase; a macroemulsion is a dispersion of droplets or particles that contains two or more phases, which are liquids or liquid crystals (48).

Nevertheless, possibilities for confusion abound. From the definitions of microemulsions and macroemulsions and from Figure 1, it immediately follows that in many macroemulsions one of the two or three phases is a microemulsion. Until recently (49), it was thought that all nonmultiple emulsions were either oil-in-water (O/W) or water-in-oil (W/O). However, the phase diagram of Figure 1 makes clear that there are six nonmultiple, two-phase morphologies, of which four contain a microemulsion phase. These six two-phase morphologies are oleic-in-aqueous (OL/AQ, or O/W) and aqueous-in-oleic (AQ/OL, or W/O), but also, oleic-in-microemulsion (OL/MI), microemulsion-in-oleic (MI/OL), aqueous-in-microemulsion (AQ/MI), and microemulsion-in-aqueous (MI/AQ) (49).

Although they have not yet all been reported, theoretically there are twelve three-phase emulsion morphologies formed by the top, microemulsion (ie, middle), and bottom phases (50,51): totally engulfing $T/M/B$, $M/T/B$, $M/B/T$, $B/M/T$, $B/T/M$, $T/B/M$; nonengulfing $(T+M)/B$, $(B+M)/T$, and $(B+T)/M$; and partially engulfing $(T+M)/B$, $(B+M)/T$, and $(B+T)/M$. These twelve three-phase emulsion morphologies, in which one of the phases is a middle-phase microemulsion, are illustrated by Figure 7.

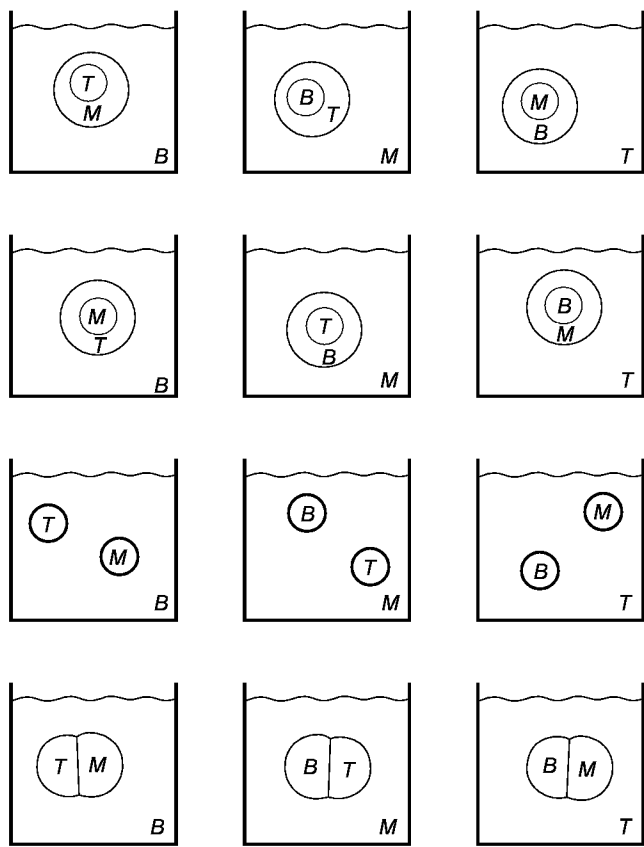


Fig. 7. Cartoon illustration of the twelve theoretical morphologies of three-phase macroemulsions in which one the phases is a middle-phase microemulsion.

Emulsion Morphology Diagrams

Just as phase diagrams clarify what constitutes a microemulsion, emulsion morphology diagrams (52–54) provide maps that clarify the relationships among microemulsions and the various emulsion morphologies in which these microemulsions can appear. The emulsion morphology diagram of Figure 8 illustrates each of the six two-phase emulsion morphology regions and, for the various regions within the tietriangle, shows which phase is the continuous phase. Separating the different morphology regions (eg, OL MI and MI OL) are narrow hysteresis regions, within which, depending on its history, the emulsion may be in either (eg, OL MI or MI OL) morphology. For two-phase microemulsion emulsions, the hysteresis region forms a cusp that terminates at a plait point. For the *T*-continuous, *B*-continuous, and *M*-continuous regions the morphology (nonengulfing, partially engulfing, or totally engulfing) of the dispersed phases is believed (50,55) to depend on the wettability condition (56) among the three phases.

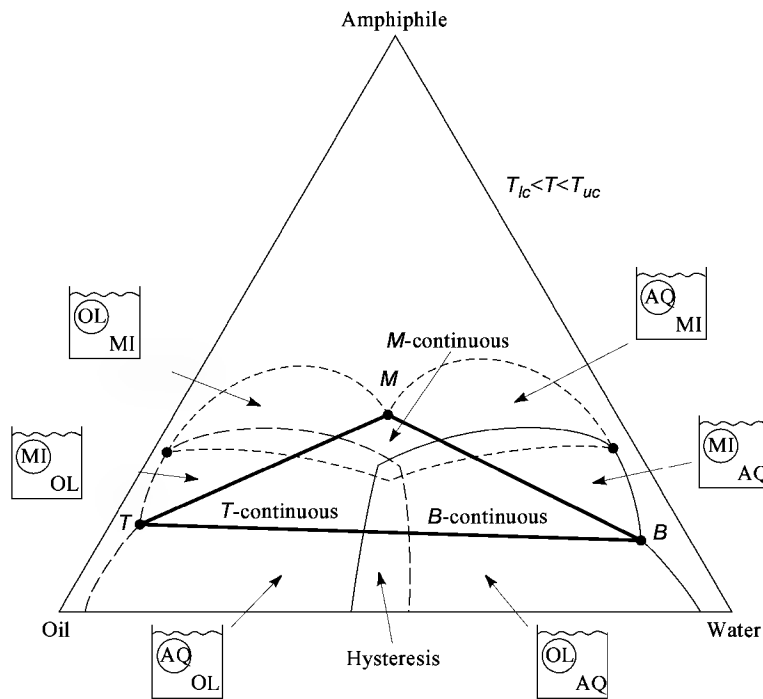


Fig. 8. Emulsion morphology diagram, illustrating where the microemulsion in various macroemulsion morphologies is a continuous phase or dispersed phase. Morphology boundaries: (), aqueous, continuous; (), oleic, continuous; (), microemulsion, continuous.

Economic Aspects

There are no statistics available for microemulsion products or their annual values, but data for the surfactant industry can be taken as a guide. Annually updated lists of commercial surfactants and their suppliers are available from several sources (57,58). *Chemical & Engineering News* annually publishes a feature article on "Soaps and Detergents" in its fourth issue (59,60). The market for surfactants is immense. For example, in 1995, the U.S. market for

laundry and dishwashing detergents, soaps, and household cleaners was \$9.6 billion (59). In 1994 the United States consumed 2.18 billion kg of anionic surfactants, 0.86 billion kg of nonionic surfactants, 0.32 billion kg of cationics, and 0.024 billion kg of amphoteric surfactants, for a total of almost 3.4 billion kg. During the period from 1994 to year 2000, the annual growths in these demands were expected to be 2.6^o %, 2.9^o %, 3.5^o %, and 6.0^o % respectively. Hence, the projected demands for the year 2000 were 2.54 billion kg for anionics, 1.02 billion kg for nonionics, 0.39 billion kg for cationics, and 0.03 billion kg for amphoterics (59). Demand for many surfactants was expected to grow even faster in other parts of the world than in the U.S. For example, in 1995, North America consumed 13^o %, Western Europe 15^o %, Latin America 19^o %, and the Asia-Pacific region 37^o % of the 2.8 million tons of linear alkylbenzene surfactants consumed in the world. For the year 2005, these fractions were predicted to be 9^o %, 8^o %, 17^o %, and 45^o %, respectively, of a total worldwide consumption of 4.0 million tons (59) (see SURFACTANTS).

BIBLIOGRAPHY

1. I. D. Robb, *Microemulsions*, Plenum Press, New York, 1982.

2. V. Degiorgio and M. Corti, *Physics of Amphiphiles: Micelles, Vesicles, and Microemulsions*, Elsevier Science Publishing Co., New York, 1985.

3. D. O. Shah, *Macro and Microemulsions: Theory and Applications*, American Chemical Society, Washington, D.C., 1985.

4. S. E. Friberg and P. Bothorel, *Microemulsions: Structure and Dynamics*, CRC Press, Boca Raton, Fla., 1987.

5. H. L. Rosano and M. Clause, eds., *Microemulsion Systems*, Marcel Dekker, New York, 1987.

6. M. Bourrel and R. S. Schechter, *Microemulsions and Related Systems: Formulation, Solvency, and Physical Properties*, Marcel Dekker, New York, 1988.

7. S. H. Chen and R. Rajagoplan, *Micellar Solutions and Microemulsions*, Springer-Verlag, New York, 1990.

8. D. Roux, *Micelles, Membranes, Microemulsions, and Monolayers*, Springer-Verlag, New York, 1994.

9. C. Solans and H. Kunieda, eds., *Industrial Applications of Microemulsions*, Marcel Dekker, New York, 1996.

10. *Global Books in Print (CD-ROM)*, R. R. Bowker, New Providence, N.J., March 1997.

11. D. H. Smith in D. H. Smith, ed., *Surfactant-Based Mobility Control: Progress in Miscible-Flood Enhanced Oil Recovery*, American Chemical Society, Washington, D.C., 1988.

12. D. R. Karsa, ed., *Industrial Applications of Surfactants*, Royal Institute of Chemistry, Cambridge, 1990.

13. M. Gratzel and K. Kalyanasundaram, eds., *Kinetics and Catalysis in Microheterogenous Systems*, Marcel Dekker, New York, 1991.

14. D. R. Karsa, ed., *Industrial Applications of Surfactants*, Royal Institute of Chemistry, Cambridge, 1992.

15. D. A. Sabatini and R. C. Knox, eds., *Transport and Remediation of Subsurface Contaminants*, American Chemical Society, Washington, D.C., 1992.

16. T. F. Tadros, ed., *Surfactants in Agrochemicals*, Marcel Dekker, New York, 1992.

17. N. Kosanic, ed., *Biosurfactants: Production, Properties, Applications*, Marcel Dekker, New York, 1993.

18. M. M. Reiger and L. D. Rhein, eds., *Surfactants in Cosmetics*, Marcel Dekker, New York, 1997.

19. S. S. Talmage, ed., *Environmental and Human Safety of Major Surfactants*, Lewis Publishers, Boca Raton, Fla., 1994.

20. D. R. Karsa and M. R. Porter, eds., *Biodegradability of Surfactants*, Blackie Academic and Professional, London, 1995.

21. M. J. Schwuger, ed., *Detergents in the Environment*, Marcel Dekker, New York, 1996.

22. T. P. Hoar and J. H. Schulman, *Nature* **152**, 102 (1943).

23. D. H. Smith, *J. Colloid Interface Sci.* **108**, 471 (1985).

24. S. E. Friberg and I. Lapczynska, *Prog. Colloid Polym. Sci.* **56**, 16 (1975).

25. H. Kunieda, *Bull. Chem. Soc. Jpn.* **56**, 625 (1983).

26. D. H. Smith and G. L. Covatch, *J. Colloid Interface Sci.* **170**, 112 (1995).

27. P. Ekwall in G. H. Brown, ed., *Advances in Liquid Crystals*, Academic Press, New York, 1975.

28. M. Kahlweit, R. Strey, and G. Busse, *J. Phys. Chem.* **94**, 3881 (1990).

29. G. D. Efremova and A. V. Shvarts, *Russ. J. Phys. Chem.* **40**, 486 (1966).

30. J. R. DiAndreth and M. E. Paulaitis in D. H. Smith, ed., *Surfactant-Based Mobility Control: Progress in Miscible-Flood Enhanced Oil Recovery*, American Chemical Society, Washington, D.C., 1988.

31. J. C. Lang, and B. Widom, *Physica* **81A**, 190 (1975).

32. M. Kahlweit, R. Strey, R. Schomaecker, and D. Haase, *Langmuir* **5**, 305 (1989).

33. D. H. Smith, *AOSTRA J. Rsch.* **4**, 245 (1988).

34. R. B. Griffiths, *J. Chem. Phys.* **60**, 195 (1974).

35. D. O. Shah and R. S. Schechter, eds., *Improved Oil Recovery by Surfactant and Polymer Flooding*, Academic Press, New York, 1977.

36. J. J. Taber, *Soc. Petrol. Eng. J.* **9**, 3 (1969).

37. G. L. Stegemeier in D. O. Shah and R. S. Schechter, eds., *Improved Oil Recovery by Surfactant and Polymer Flooding*, Academic Press, New York, 1977.

38. H. Kunieda and K. Shinoda, *Bull. Chem. Soc. Jpn.* **55**, 1777 (1982).

39. P. D. Fleming, J. E. Vinatieri, and G. R. Glinsmann, *J. Phys. Chem.* **84**, 1526 (1980).

40. B. Widom, *J. Chem. Phys.* **62**, 1332 (1975).

41. B. Widom, *Phys. Rev. Lett.* **34**, 999 (1975).

42. W. C. Griffin, *J. Soc. Cosmetic Chemists* **1**, 311 (1949).

43. H. Kunieda and K. Shinoda, *J. Disp. Sci. Technol.* **3**, 233 (1982).

44. K. Shinoda and H. Kunieda in P. Becher, ed., *Encyclopedia of Emulsion Technology*, Vol. 1, Marcel Dekker, New York, 1983.

45. J. S. Rowlinson and B. Widom, *Molecular Theory of Capillarity*, Clarendon Press, Oxford, 1984.

46. P. D. Fleming and J. E. Vinatieri, *J. Colloid Interface Sci.* **81**, 319 (1981).

47. D. H. Smith, *J. Chem. Phys.* **85**, 1545 (1986).

48. *International Union of Pure and Applied Chemistry Manual on Colloid and Surface Science*, Butterworths, London, 1972.

49. D. H. Smith in H. L. Rosano and M. Clause, eds., *Microemulsion Systems*, Marcel Dekker, New York, 1987.

50. D. H. Smith, G. K. Johnson, Y. C. Wang, and K.-H. Lim, *Langmuir* **10**, 2516 (1994).

51. G. K. Johnson, D. Dadyburjor, and D. H. Smith, *Langmuir* **10**, 2523 (1994).

52. D. H. Smith and K.-H. Lim, *J. Phys. Chem.* **94**, 3746 (1990).

53. D. H. Smith, J. S. Reckley, and G. K. Johnson, *J. Colloid Interface Sci.* **151**, 383 (1992).

54. D. H. Smith and Y.-C. Wang, *J. Phys. Chem.* **98**, 7214 (1994).

55. S. Torza and S. G. Mason, *Science* **163**, 813 (1969).

56. D. H. Smith and G. L. Covatch, *J. Chem. Phys.* **93**, 6870 (1990).

57. *Chemyclopedia* (Annual Suppl. to *C&E News*), American Chemical Society, Washington, D.C., 1997.

58. *McCutcheon's Emulsifiers & Detergents, International Ed.*, MC Publishing Co., Glen Rock, N.J., 1997.

59. *C&E News* **74**(4), 32 (1996).

60. *C&E News* **75**(4), 30 (1997).

Duane H. Smith
Technical Solutions and West Virginia University

MOLECULAR MODELING

It can be said that science is the art of building models to explain observations and predict new ones. Chemistry, as the central science, utilizes models in virtually every aspect of the discipline. From the first week of a first chemistry course, students use the scientific method to develop models which explain the behavior of the elements. Anyone who studies or uses chemistry has, in fact, practiced some form of molecular modeling.

A useful method of tracing the origins of molecular modeling is to examine how its offshoot, computer-assisted molecular modeling (CAMM), came to be developed. Molecular modeling today represents a convergence of a number of techniques from different disciplines. Basic techniques used to accomplish modeling objectives necessarily draw on these. Specific software systems today importantly assist the researcher in the study of molecular systems and provide help in deriving a rigorous and consistent explanation for the chemical or biological behavior observed or help the researcher to develop a model for predictions.

The literature of molecular modeling is expanding rapidly. The reasons the number of molecular modeling studies continues to grow are that the tools are becoming easier to use, computers get faster, and researchers' awareness of the utility of these tools is growing. Despite this increasing ease to use, however, the development of a good grasp of the origins of the methods and strategies of molecular modeling could usefully draw on the many citations from the literature that are included herein.

Historical Perspective

Molecular modeling has evolved as a synthesis of techniques from a number of disciplines—organic chemistry, medicinal chemistry, physical chemistry, chemical physics, computer science, mathematics, and statistics. With the development of quantum mechanics (1,2) in the early 1900s, the laws of physics necessary to relate molecular electronic structure to observable properties were defined. In a confluence of related developments, engineering and the national defense both played roles in the development of computing machinery itself in the United States (3). This evolution had a direct impact on computing in chemistry, as the newly developed devices could be applied to problems in chemistry, permitting solutions to problems previously considered intractable.

Into the late 1940s, Nobel Laureate Robert S. Mulliken, a physical chemist at the University of Chicago, maintained a skeptical view regarding the future of applying the theories of physics to solving practical problems in chemistry (4,5). Subsequently, Mulliken (5) related that

... [it was] only in the '50s that really substantial progress was made ... A major and indeed crucial step beyond the development of formulas for molecular integrals was the programming for large electronic digital computers of otherwise excessively time-consuming numerical computation of these integrals, and of their combination to obtain desired molecular wave functions and related molecular properties.

Whereas many scientists shared Mulliken's initial skepticism regarding the practical role of theory in solving problems in chemistry and physics, the work of London (6) on dispersion forces in 1930 and Hückel's π -electron theory in 1931 (7) continued to attract the interest of many, including a young scientist named Frank Westheimer who, drawing on the physics of internal motions as detailed by Pitzer (8), first applied the basic concepts of what is now called molecular mechanics to compute the rates of the racemization of *ortho*-dibromobiphenyls. The 1946 publication (9) of these results would lay the foundation for Westheimer's own systematic conformational analysis studies (10) as well as for many others, eg, Hendrickson's (11) and Allinger's (12). These scientists would utilize basic Newtonian mechanics coupled with concepts from spectroscopy (13,14) to develop nonquantum mechanical models of structures, energies, and reactivity.

Researchers in chemistry and chemical physics whose interests were more theoretical were also very active between 1940 and 1965. The Manhattan Project had focused a great deal of public attention on chemistry and physics, garnering as well the energy and interest of both young and accomplished scientists. Coulson's seminal contribution (15) on molecular orbitals, along with studies of the low lying excited states of benzene by Goeppert-Mayer and Sklar (16), laid the groundwork for the subsequent contributions of Dewar (17), Pople (18), and Pariser and Parr (19), whose bodies of work have provided the basis for innumerable citations detailing the application of quantum mechanical methods to organic chemistry since the 1950s, a subject thoroughly treated in several exemplary texts (20–22). Such efforts (17–19) paved the way for Nobel Laureates Woodward and Hoffman (23) to elaborate elegant orbital-based theories on the relationship between reactivity and molecular electronic structure.

Definition of Molecular Modeling and Uses of CAMM

Molecular modeling refers broadly to any study of molecules utilizing physical or theoretical models to explain an observed or predicted behavior. In practice, physical models have expedited the understanding of small molecules, inorganic complexes, and biopolymers, including such molecules as DNA. The pioneering work of Watson and Crick on DNA would certainly not have been possible without the building of actual physical models of the structure derived from experimental observations of x-ray diffraction patterns. However, physical models have limitations, both in terms of being primarily static representations of dynamic systems, and in being only semiquantitative with respect to scale. Whereas chemists prefer having hands-on experience with actual physical models, their parallel quest to quantify structures and energetics makes it impractical to rely only on such models for all aspects of their work. It is here that computer models, or molecular modeling, enter in.

Molecular modeling can be defined as the application of computational techniques, grounded in theory, to predict or explain observable biological or physical chemical properties. Wherever molecular modeling is practiced using a computer, the technique then becomes computer-assisted (aided) molecular modeling, or CAMM. CAMM, is often used synonymously with CAMD, or computer-assisted molecular (materials) design discovery. CADD refers to computer-assisted drug design discovery. A computational technique as used herein is a mathematical model derived from principles of chemistry, physics, or statistics which facilitates molecular modeling. An entire branch of chemistry, ie, computational chemistry, is devoted to developing, benchmarking, and applying computational techniques in order that researchers may be able to better understand and predict properties. Computational chemistry serves as an umbrella under which several disciplines converge to promote the evolution of better technologies to enhance the understanding of molecules and their reactivities. Some of the properties which may be calculated either exactly or approximately by computational methods are the following:

- boiling points
- melting points
- crystallization energy
- heat capacity
- heat of formation
- heat of fusion

heat of sublimation
heat of vaporization
entropy
molar refractivity
molar volume
partition coefficients
radius of gyration
elasticity
tensile strength
dipole moments
quadrupole moments
octupole moments
infrared spectra intensities
nmr spectra chemical shifts
optical rotary dispersion
raman spectra
ultraviolet spectra
ionization potentials
electron affinities
protonation energies and pK_a s
ionic strength
conformational energies
Boltzmann distributions

Molecular properties can be classified according to their end-point observables, such as chemical (reactivity, solubility, acid–base), physical (a function of physical state—gas, liquid, solid; thermodynamic), or biological (ligand or enzyme; agonist or antagonist). These properties reflect macroscopic, or bulk, properties, which exist only for the bulk material, eg, heat of crystallization, or microscopic properties, which exist for an ensemble of the molecule. As use of CAMM methods expands to address a broader horizon of applications beyond those in organic, medicinal, and biological chemistry, calculations on metals, semiconductors, and magnetic systems are becoming more common (24).

To gain a proper perspective of the role of computed physical properties, the relationship between estimated and computed properties needs to be understood. A thorough reading of Horvath (25) permits formulation of the following definitions of estimating or computing properties.

Interpolating properties: a correlation is found between the desired property and another property or characteristic of related molecules; in this case the desired property may be computed from within the range of application of the correlation, and the interpolated property should be accurately estimated.

Extrapolating properties: in this case the correlation does not extend to include the molecule of interest, but by extending the correlation, it is possible to estimate the desired property. Since the validity of the correlation in the extrapolated region is unknown, the accuracy of the extrapolated property is difficult to estimate.

Computing properties: in many cases it is possible to compute a property, directly or indirectly, with varying levels of accuracy. Such computed properties can be quite comparable to experimental accuracy and indeed may substitute for the experiment in cases where the experiment would be difficult or impossible to perform.

This last definition should be carefully applied as either an interpolation or an extrapolation, particularly for empirical computational methods based on diverse observations. It is critical that users of molecular modeling tools understand where it is appropriate to apply a technique and where it is not, and what degree of accuracy can be expected.

The specific process involved in a given molecular modeling study depends significantly on the nature of the primary objective of the task. However, for many small-molecule and even macromolecular studies, a number of authors have diagrammed the individual steps in the process, and the flow chart in Figure 1 summarizes their efforts. An initial step that is critical to any CAMM project is the generation or retrieval of the pertinent structures themselves, a subject which is treated in detail herein. Structures may range from simple organic molecules or monomers of more complex polymers, to full proteins, enzymes, metal surfaces, or zeolites. The modeling process can be influenced by the initial structure and its geometry. Thus, the selection and development of the starting molecular geometry needs to be given particular attention. An important component of the quality control process, as well as in gaining an understanding of the molecules themselves, is the visual examination of structures involved in a modeling study.

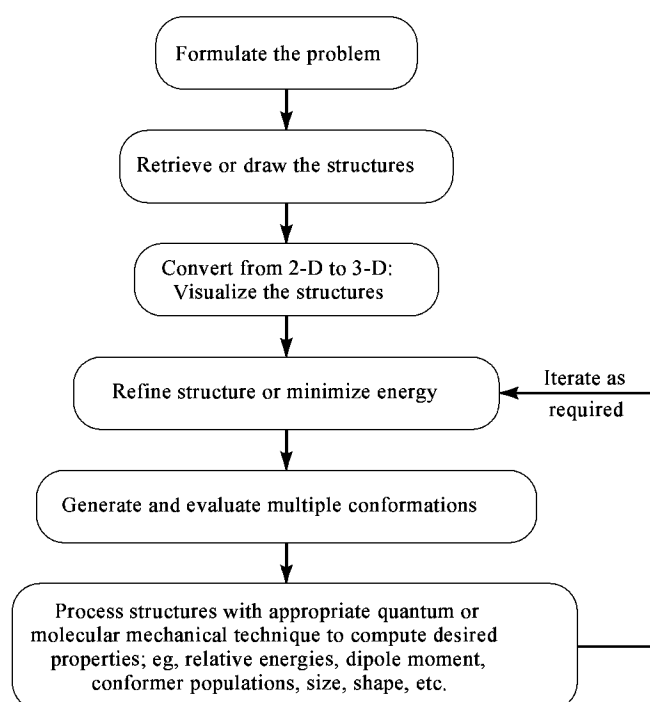


Fig. 1. Flow chart for a typical small molecule modeling project.

Computer Graphics In Molecular Modeling

The goal of molecular modeling is to define clearly the relationship between chemical constitution, ie, the molecular formula or a topographic representation thereof, its geometric constitution or 3-D topology (the disposition of its atoms in Cartesian space), and its observed (or predicted) properties. The representation and facile manipulation of 3-D arrays of atoms comprise the domain of molecular or computer graphics. From the time of Levinthal's work in the mid-1960s (26), scientists have endeavored to develop and use graphics software and hardware to expedite molecular modeling studies in both the classroom and the research lab. The growth of graphics tools has paralleled the evolution of computing hardware and, indeed, in some circumstances has even given strong impetus to that evolution (27). An example is the development and use of the Evans and Sutherland computer graphics systems in molecular modeling (27). Computing software and hardware systems have had a profound impact on the ability of modelers to compose a modeling study and address all aspects of the work, ranging from generating 2-D drawings of structures to statistical quality control of computed properties via visualization of multidimensional data.

Illustrating this enhancement in the visualization of structure and properties, Figures 2–5 provide increasingly complex and useful structural representations. Figure 2, a stick drawing of the drug Zantac (ranitidine), when viewed in color (not shown here), conveys both 2-D structure data and color-coded electronic charge data. Figure 3 shows the CPK shaded solid surface for ranitidine. In Figure 4, color-coded electronic charge data (color not shown here) have been mapped onto the CPK surfaces. In Figure 5, four layers of the Connolly solvent-accessible dot surface, when color-coded (color not shown here), correspond to the energies of the electrostatic potential. In this figure, the highest charge density, when viewed in color, would be indicated by red dots in areas where there is the strongest attraction to an H^+ atom brought to that point. Conversely, at the points that would be shown in purple, the repulsive interaction would be the greatest, signifying regions of maximal positive charge. Figure 6 shows the electrostatic potential of ranitidine mapped, respectively, onto a solid and translucent constant-density surface ($0.001 e$). Figure 6 is actually five-dimensional when viewed in color (x , y , and z dimensions, augmented by the color and gradients of color not shown here). The added impact of color ordinarily shown in such images considerably augments what the reader is able to view in black-and-white figures. An excellent overview of computer graphics by Hubbard (28) provides details regarding the mechanics of hardware and software systems integral to current (1996–1997) methods of computer graphics. A continuing source of papers on methods and applications in computer graphics and molecular modeling can be found in the *Journal of Molecular Graphics and Modeling* (Elsevier).

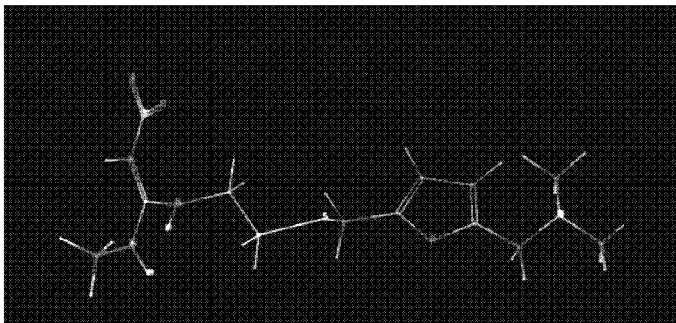


Fig. 2. A stick drawing of the drug Zantac (ranitidine) illustrating the method of showing color-coded electronic charge data (color not shown here), where red would represent the highest charge density, blue the lowest charge density, etc.

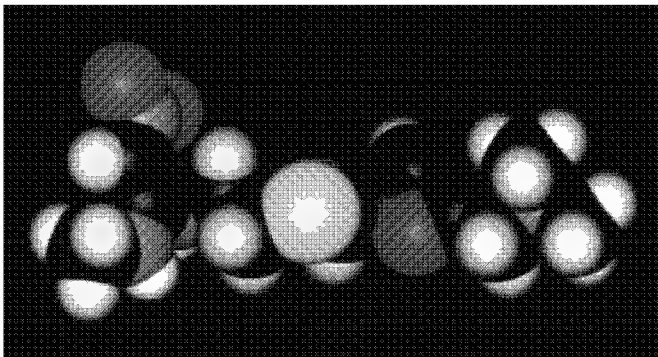


Fig. 3. CPK (Corey-Pauling-Koulton) rendering displaying shaded solid surface for ranitidine, normally color-coded according to atom types (black representing carbon; green, hydrogen; yellow, sulfur; blue, nitrogen; and red, oxygen; colors not shown here).

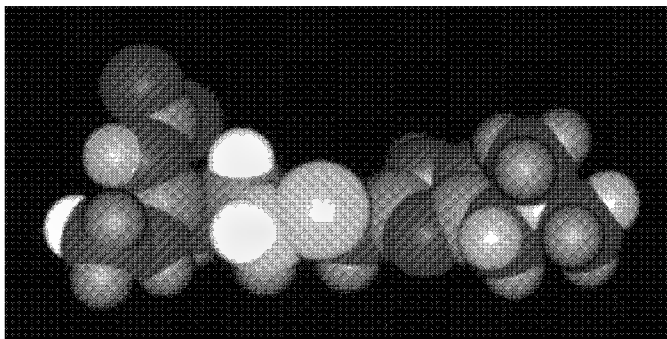


Fig. 4. CPK rendering of ranitidine as in Fig. 3, although here illustrating the method of showing color-coded electronic charge data mapped onto the CPK surfaces, with red representing the highest density, blue the lowest density, etc (color not shown here).

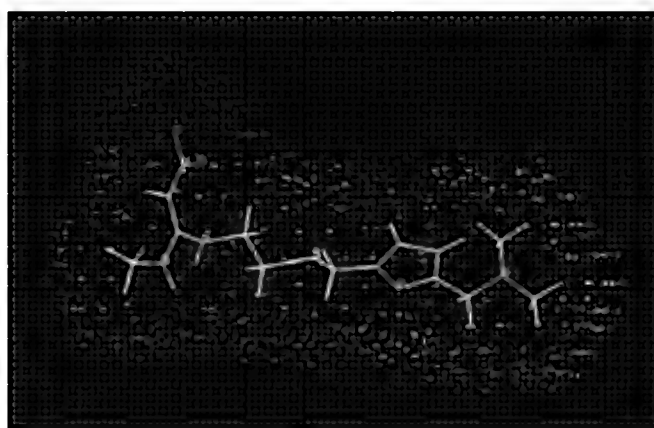


Fig. 5. A representation of ranitidine displaying four layers of the Connolly solvent-accessible dot surface normally color-coded in this process to correspond with the energies of electrostatic potential (color not shown here). Thus, the highest charge density would be indicated by red dots representing points where the attraction to an H^+ atom is strongest, and conversely, purple points would signify regions of maximal positive charge.

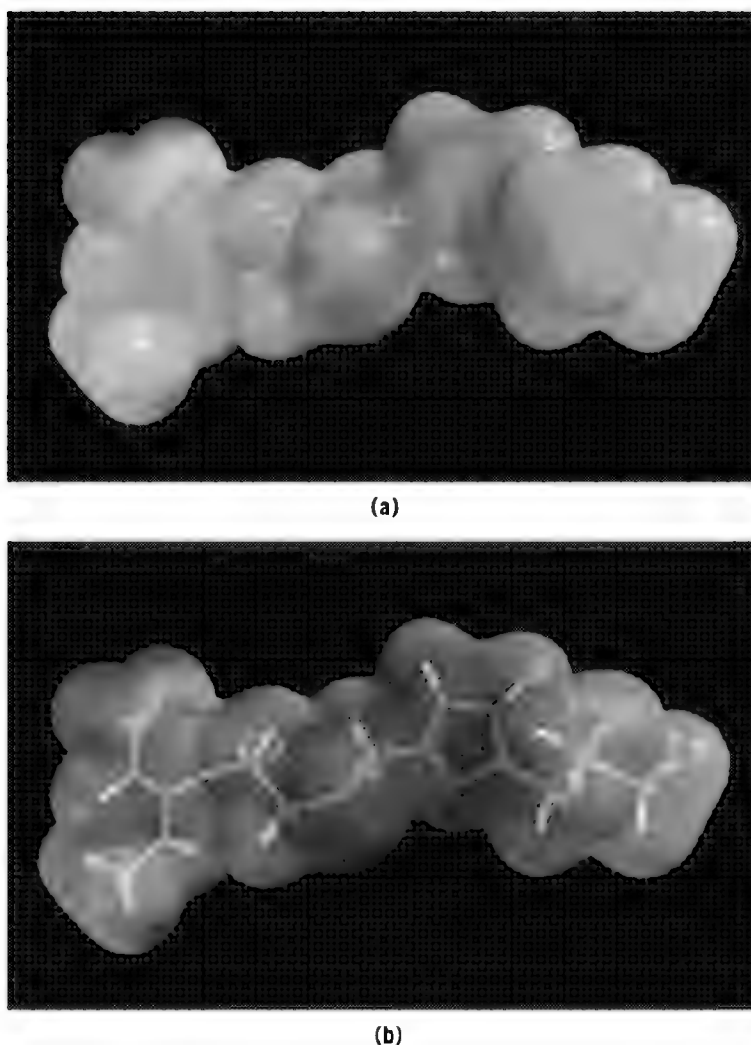


Fig. 6. Two representations of ranitidine illustrating the electrostatic potential mapped onto (a) a solid and (b) translucent constant-density surface (0.001 e); see text for effect of color-coding, not shown here.

One facet of chemistry which has benefited greatly from the use of computer graphics to enhance its own development is x-ray crystallography; the use of computers in x-ray structure refinement has also been described (29). Reviews (30) have appeared which cover the development and range of computing methods as adjuncts to crystallographic studies. More recently, details of how computer graphics can provide visual introductions to basic diffraction concepts, including the reciprocal lattice, the Ewald sphere construction, lattices, space group determination, and many other topics, have been published (31). One of the most memorable advances in graphics in modeling was showcased on the cover of *Science* magazine in 1981. Researchers (32) in pharmaceutical and physical chemistry as well as computer and information science at the University of California at San Francisco's Laboratory for Computer Graphics worked together to produce a highly symmetrical graphical representation of the B-DNA molecule viewed along its helix axis. The view produces the effect of a stained-glass window. Although this effect was impressive in both a scientific and an artistic sense, it was the representation of the molecular surface as a series of dots following the work of Connolly (33), which provided the real advance in molecular graphics; a review of molecular surfaces is given in Reference 34. Subsequently, a number of computed properties such as the molecular electrostatic potential (see Fig. 5) and the relative hydrophobicity have been mapped onto the Connolly surface, which has become an indispensable part of molecular modeling in the drug industry.

Among the current (ca 1997) selection of software systems which together help to exemplify the promise of computer graphics, Advanced Visualization Systems (AVS), stands out as one of the more extensible and practical of them to use. The premise behind development of the system was to provide modelers with a toolkit of modules having sophisticated intrinsics that would enable even casual programmers to link together multiple simple functionalities into a complex construct with which to accomplish exactly the types of visualization and manipulations that their work required. Researchers

at the North Carolina Supercomputing Center helped to assemble a vast library of modules which are available on the Internet (<http://www.avsc.com>), and they have published work on the utilization of AVS in molecular modeling which describes methods for combining orbital descriptions of molecules, as well as models of how the AVS visualization utilities are applied to the understanding of such properties as the electrostatic potential and reactivity indices (35). The utilization of computer graphics tools has also been extended to viewing the chemical literature itself (36) through the introduction of programs and data sources for "kinemages," ie, kinetic images that can be viewed and rotated interactively on a PC or Macintosh. Such capabilities are now more common through utilization of the Internet and via plug-ins to Web browsers.

Computational Methods for Molecular Modeling

Classical and Quantum Mechanics. At the beginning of the twentieth century, a revolution was brewing in the world of physics. For hundreds of years, the Newtonian laws of mechanics had satisfactorily provided explanations and supported experimental observations in the physical sciences. However, the experimentalists of the nineteenth century had begun delving into the world of matter at an atomic level. This led to unsatisfactory explanations of the observed patterns of behavior of electricity, light, and matter, and it was these inconsistencies which led Bohr, Compton, deBroglie, Einstein, Planck, and Schrödinger to seek a new order, another level of theory, ie, quantum theory.

Basically, Newtonian mechanics worked well for problems involving terrestrial and even celestial bodies, providing rational and quantifiable relationships between mass, velocity, acceleration, and force. However, in the realm of optics and electricity, numerous observations seemed to defy Newtonian laws. Phenomena such as diffraction and interference could only be explained if light had both particle and wave properties. Indeed, particles such as electrons and x-rays appeared to have both discrete energy states and momentum, properties similar to those of light. None of the classical, or Newtonian, laws could account for such behavior, and such inadequacies led scientists to search for new concepts in the consideration of the nature of reality.

In 1903, when Planck suggested that the energy emitted from heated bodies, ie, black-body radiation, was not composed of waves, but rather discrete particles or quanta, a long-standing physical anomaly was resolved. Similarly, Planck's theory was applied to the photoelectric effect and was subsequently used by Bohr (37) to develop models of atomic structure. By the time Pauling and Wilson published their treatise on quantum mechanics in 1935 (38), the foundations for a workable quantum theory, explaining black-body radiation, electron distributions around nuclei and in chemical bonds, and the wave-particle duality of photons and electrons, had been detailed by a new generation of physicists (39). The extraordinary progress in the theory of matter made during the first three decades of the twentieth century lead Dirac, one of the pioneers of quantum theory 40, to state, "The underying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known" (41).

At the heart of the revolution in quantum theory is Schrödinger's equation, which, in one dimension, for one electron not interacting with its surroundings, may be written

$$\frac{d^2\Psi}{dx^2} + 8\pi^2m \cdot h^2(E - V) \Psi = 0$$

in which E is the total energy, a constant, and V is the potential energy, which most often is a function of x . The wave function, Ψ , provides the solution of this equation and is at the heart of quantum mechanics and its application to problems in chemistry. After expanding the wave equation to three dimensions, and replacing the second derivatives in equation 1 with the Laplacian operator, ∇^2 , equation 1 can be rearranged to give

$$\mathbf{H} = T + V$$

wherein the energy terms have been on grouped the right, and the Hamiltonian operator, $\mathbf{H}$, is used in classical mechanics to provide a function of momenta and coordinates. For a single electron, the Schrödinger equation is

$$\mathbf{H}\Psi = E\Psi$$

in which the $\mathbf{H}$ is an operator, the constant E is called an eigenvalue, and the function Ψ is called the eigenfunction. Certain constraints must be met if Ψ is to be physically meaningful, ie, that it be continuous and single-valued over the region of interest. The probability of finding the particle in all space must be unity, ie,

$$\int |\Psi|^2 dx dy dz = 1$$

The quantity $|\Psi|^2$ represents the probability of finding the electron in the described region. It is also interpreted statistically, in the context of the Heisenberg uncertainty principle, as an expectation value. Streitwieser has eloquently summarized the importance of such concepts in chemical computations (42):

For organic chemists the importance of quantum mechanics lies not at all in the exact calculations from first principles (*ab initio* calculations), but rather in providing heuristic concepts and insights in establishing qualitative and quantitative semiempirical correlations of experimental data and, especially, in facilitating the application of what has long been the organic chemist's most important tool: reasoning by analogy.

Two important points are made in this statement: first, less important than exact quantification is the development of a heuristic model of chemical behavior; and second, semiempirical correlations are the goal of computations involving applied quantum mechanics. At the time Streitwieser wrote, in 1960–1961, fullscale calculations on systems large enough to be of interest to organic chemists were certainly beyond then-current limits for most levels of theory and the corresponding computer programs. Computing machinery was still quite primitive by today's standards (about 10 slower than a good workstation at the time of this writing, ie, early 1997), and a full elaboration of semiempirical quantum mechanical methods was in progress (20). Of particular note is the fact that Streitwieser here suggests that organic chemists are less interested in exact *ab initio* calculation of properties than in developing qualitative concepts of structure–reactivity relationships. As discussed herein, *ab initio* quantum mechanical calculations now have taken their place alongside semiempirical methods in the chemist's computational toolkit for molecular modeling.

Clearly, most chemistry involves interactions of multielectron systems with each other to yield either desired quantities of previously characterized chemicals or new chemical entities (NCEs, in pharmaceutical language). At the atomic level, chemical reactions involve structures which have multiple electrons interacting, perhaps even on multiple atomic centers such that the potential energy of these systems is dependent on the relative positions of the electrons with respect to each other as well as with respect to the nuclei. The Born-Oppenheimer approximation is generally applied; this approximation states that the motion of the nuclei is neglected, and that electron motion is the dominant component of interatomic interactions. Hartree (43) and Fock (44) developed a formalism for reducing the multielectron Schrödinger equation to a sum of single-electron equations, which could be solved to yield what is called a Self-Consistent Field (SCF) Approximate Method. As an interactive method wherein electrons are distributed into shells based on the Aufbau principle, the SCF method made it possible to develop wavefunctions which can be approximated by analytical solutions to provide representations of atomic and molecular orbitals that can be readily visualized. Such visualization facilitates the interpretation of chemical reactivity and chemical reactions.

Numerous methods arose utilizing Hartree-Fock SCF techniques, ranging from the simplest, or Hückel π -electron techniques, to the most complete

first principles, or *ab initio* methods. What distinguishes these methods is, in practical terms, which electrons and orbitals are included in the calculations, along with the degree to which the elements of the Fock matrix (representing the operator $\mathbf{H}$) are evaluated explicitly, approximated, or neglected completely. These matrix elements, when they are evaluated in the context of the linear combination of atomic orbitals (LCAO) method, ultimately provide a quantitation of the presence and magnitude of the influence of neighboring electrons on each other. For systems in which all σ - and π -electrons are considered, so long as the researcher neglected a selected set of neighboring atom electron–electron interactions LCAO provided the basis for most so-called semiempirical quantum mechanical methods in use around 1960.

The Hückel Molecular Orbital theory (HMO) and its subsequent elaboration, Extended HMO theory (EHT), methods provide the simplest quantum mechanical description of π -electron systems. The development and applications of HMO were reviewed by Streitwieser in his text (20). Whereas the HMO method neglects σ -electrons, it did utilize linear combinations of atomic orbitals (LCAO) for the molecular orbitals. The method worked quite well, particularly for planar conjugated systems and even for certain nonplanar molecules. However, the technique can be best credited as providing the basis for subsequent, more elaborate methods. In Murrell and Harget (22) the development and applications of the Piser-Pople-Parr (PPP) (19) method is given in a very readable account. It was a particularly useful technique in that adiabatic ionization potentials as well as singlet-triplet separation energies could be computed accurately. Additionally, the PPP method was used subsequently to compute the energetics of the π -electron component of structures (45,46), in conjunction with an adaptation of the Westheimer method to compute gas-phase conformational energies for planar and nonplanar π -electron systems.

The next step in the development of molecular orbital theories involved all-valence-electron methods wherein the concept of zero differential overlap (ZDO) of two-center integrals involved in the wave functions are refined. In these techniques, which introduced drastically reduced numbers of integrals requiring evaluation, investigators found they could incrementally move toward a more complete set of SCF equations without their being computationally intractable. In 1965, the first group of a series of important papers detailing the Complete Neglect of Differential Overlap (CNDO) and Neglect of Diatomic Differential Overlap (NDDO) methods were published (47). Full computational details of these methods are also available (21,22). The CNDO and NDDO techniques enabled computation of a broad spectrum of geometric features, such as bond lengths, angles, and related properties, including dipole moments, which could be predicted for singly bonded systems for the first time. The methods continue to be used today (ca 1997), although the relatively poor accuracy of the CNDO technique in determining structural and charge distribution properties has led to its being used principally in spectroscopy (48).

Through the 1970s and 1980s investigators continued to improve on the ZDO technique, especially in increasing the accuracy and range of computed quantities, and in taking advantage of the enhanced computing facilities which were becoming available. Advances made included development of a series of programs utilizing the Intermediate Neglect of Differential Overlap (INDO) technique (49). The "Dewar School" continues today to produce semiempirical techniques to permit the computation of properties of organic molecules, organometallics, semiconductors, peptides, and proteins (50,51). Excellent expositions on semiempirical methods derived from NDDO are available (51,52), including a "how to" text (53). However, as Dewar has stated, "MO (Molecular Orbital) Theory is not a description of reality. It is only the embodiment of another molecular model, the MO model" (58).

The ultimate goal of quantum mechanical calculations as applied in molecular modeling is the *a priori* computation of properties of molecules with the highest possible accuracy (rivaling experiment), but utilizing the fewest approximations in the description of the wave function. *Ab initio*, or from first principles, calculations represent the current state of the art in this domain. They are also referred to as nonempirical calculations, although this name is somewhat misleading. *Ab initio* calculations utilize experimental data on atomic systems to facilitate the adjustment of parameters such as the exponents of the Gaussian functions used to describe orbitals within the formalism. Additionally, these Gaussians are a function of the electron–nucleus distance squared, r^2 , and represent a significant approximation to the Slater-type or simple exponential functions of r , used to describe electron distributions in elemental quantum mechanics. Whereas Gaussian functions do a creditable job of reproducing experimental properties, they were introduced for pragmatic reasons, namely, to simplify the computation of multicenter integrals. A full and readable exposition and detailing of *ab initio* molecular orbital calculations, including the very important topic of basis sets, is available (54).

The performance of *ab initio* techniques distinguishes them significantly from their predecessors, semiempirical methods. Their consistent reproduction of data from structural, thermodynamic, and reaction sources to a range falling within the error limits of the experimental values provides scientists with an important tool with which to address various modeling problems. Whereas the absolute value of the relative performance of *ab initio* techniques varies for each structural or energetic feature examined, it is not unreasonable to suggest that if the quantity can be computed by both semiempirical and *ab initio* methods, the *ab initio* value will be closer to experiment or an ideal value than any other method. To get a fuller appreciation of the comparison of these two methods, one need only follow the numerous publications of the Pople (55) group detailing the performance of *ab initio* methods in comparison to these published by the Dewar (49) group, which continued the refinement of semiempirical techniques. A mathematically rigorous overview of the utility of *ab initio* calculations is available (56), and many elementary questions about the role of *ab initio* calculations in molecular modeling studies are addressed in a particularly readable account (57) on the basis of which several important points about both semiempirical and *ab initio* calculations are worth summarizing herein.

Ab initio Theory: Caveats and Performance.

Except for approximations made previously, no others are made.

All electrons are treated.

All electron integrals are computed exactly, but with no guarantee of accuracy of the prediction or agreement with experiment.

Basis sets consist of a finite number of Gaussian functions, introducing an inherent limitation on accuracy.

Computing time can be proportional to the fourth power of the number n of basis functions.

Whereas correlation energies can be included, in practice it is even more time-consuming to use them in the determination of molecular geometries (ie, n), and determining the correct basis set to use can be difficult.

Geometric properties are quite sensitive to the basis set chosen, including the presence or absence of polarization functions (additional s and p -type functions on H and d -type on heavy atoms).

Geometric and energetic properties are also sensitive to the starting geometries, and to the algorithm used for geometry optimization.

The following conclusions apply to organic molecules of about 25 heavy atoms (~60 atoms total), assuming use of medium-size basis sets (3–21G⁺):

Geometric properties can be reproduced to within 0.15 ± 0.15 nm (0.015 ± 0.015 Å) for bond lengths, $1\text{--}2^\circ$ for bond angles, and to $\pm 5^\circ$ for dihedral angles.

Ionization energies can be computed to about ± 0.2 eV; rotational barriers to about 0.5 kcal/mol; dipole moments to about ± 0.5 D; barriers to inversion to about ± 2.5 kcal/mol; infrared frequencies can be computed with about a 15° error (usually too high); and protonation energies are accurate to about 1 pK unit.

Hydrogen bond geometries may be reproduced or predicted fairly well with reasonable, but sometimes underestimated heavy atom–heavy atom distances; radial dependence of the hydrogen bond may be in error.

Semiempirical MO Theory: Caveats and Performance. The same basic theoretical assumptions are made as in *ab initio* theory.

Only valence electrons are considered, and the influence of core shell electrons are accommodated by a nuclear screening factor.

The total number of integrals computed depends greatly on the level of complexity of the method: time cost savings of 2 orders of magnitude can be realized in *ab initio* theory (n^2 vs n^4).

Some of the elements of the Fock matrix are set to constants which have been empirically adjusted to reproduce certain properties.

Slater-type orbitals (STOs) are used to represent electron density around an atom.

Whereas correlation energies can be included, in practice it is more time-consuming to use them in the determination of molecular geometries (ie, n^3), and

determining the correct approximate method to use can be very difficult.

Geometric properties are quite sensitive to the method chosen, including the presence or absence of polarization functions (additional *s*- and *p*-type functions on H and *d*-type on heavy atoms).

Geometric and energetic properties are also sensitive to the starting geometries, and to the algorithm used for geometry optimization.

The following conclusions apply to organic molecules of about 25 heavy atoms (~60 atoms total), assuming use of the MNDO or AM1 Hamiltonian:

Geometric properties can be reproduced to within 0.3 ± 0.2 nm (0.03 ± 0.02 Å) for bond lengths, $2\text{--}4^\circ$ for bond angles, and to $\pm 10^\circ$ for dihedral angles.

Ionization energies can be computed to about ± 0.3 eV; rotational barriers to about 1–1.5 kcal/mol; dipole moments to about ± 0.3 D; barriers to inversion to about ± 5 kcal/mol; and protonation energies are accurate to about 1–2 pK units.

Hydrogen bond geometries may not be reproduced or predicted well at all, with anomalous heavy atom–heavy atom distances, and no radial dependence of the hydrogen bond. Check that the method being used has been recalibrated for this use if needed.

Thermodynamic properties such as heats of reaction and heats of formation can be computed more reliably by *ab initio* theory than by semiempirical MO methods (55). However, the literature of the method appropriate to the study should be carefully checked before a technique is selected. Finally, the role of computer graphics in evaluating quantum mechanical properties should not be overlooked. As seen in Figures 2–6, significant information can be conveyed with stick models or various surfaces with charge properties mapped onto them. Additionally, information about orbitals, such as the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), which are important sites of reactivity in electrophilic and nucleophilic reactions, can be plotted readily. Figure 7 shows representations of the HOMO and LUMO, respectively, for the antiulcer drug Zantac.

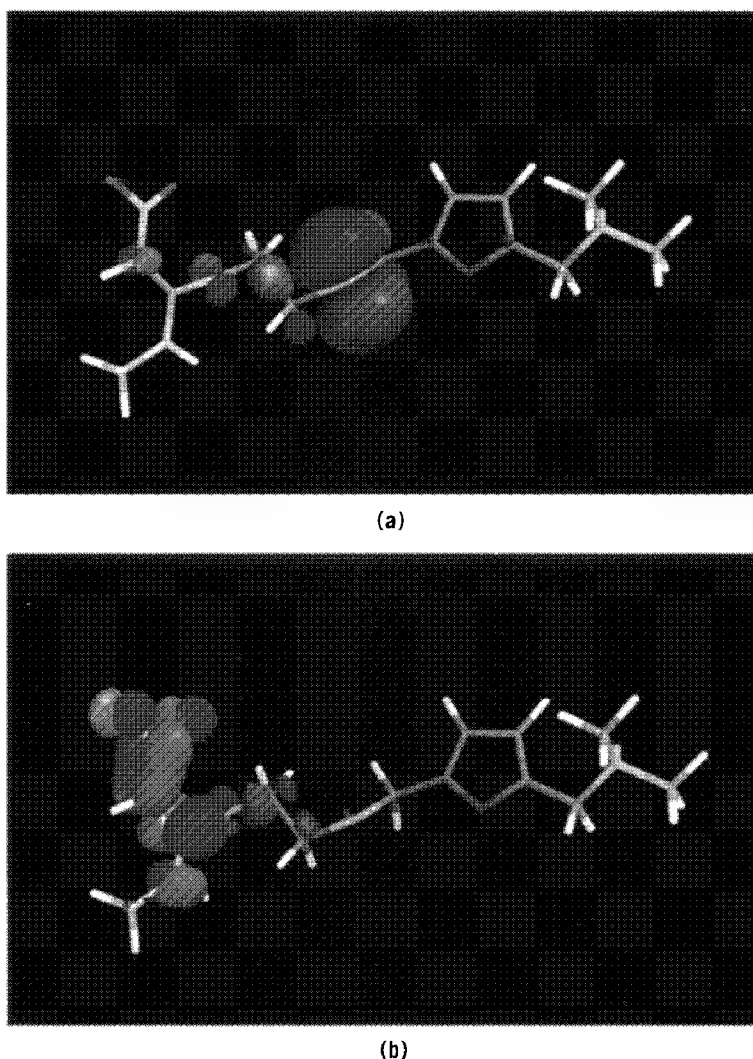


Fig. 7. Graphical representations of (a) the Highest Occupied Molecular Orbital (HOMO) and (b) the Lowest Unoccupied Molecular Orbital (LUMO) for ranitidine. It is possible, in the ordinarily visible color-coded data not shown here, to distinguish the strong localization (a) of the HOMO to the sulfur atom and adjacent nitroethylethylamine fragment and the contrasting localization (b) of the LUMO to the nitroethylethylamine fragment. Neither the LUMO nor HOMO appear to have contributions from the dimethylaminomethyl-substituted furan.

Molecular Mechanics and Molecular Dynamics.

Background. In the realm of quantum mechanics, researchers deal explicitly with electrons, with their interactions, and with their attraction to nuclei, albeit in varying degrees depending on the rigor of the method chosen to solve the particular problem involved. These techniques were somewhat limited in their application prior to the age of large-scale computers and super workstations. In 1946, Westheimer (9) could not readily simulate processes like the conformational interconversion of substituted biphenyls by either quantum or classical mechanics. However, an explicit solution to classical mechanical equations was possible by hand. Quantum mechanics dominated the world of computational chemistry and molecular modeling as it existed in 1950 until about 1975. Quietly, in the mid-1960s researchers in physical, organic, and physical organic chemistry began making incremental advances in computational tools, expanding on Westheimer's technique to study larger and perhaps more interesting molecular systems, including proteins. Computers made possible the development of Westheimer's method, also called molecular mechanics, or empirical force field methods and, ultimately, the application of these methods to a full spectrum of studies in structure and energetics.

Molecular Mechanics. Molecular mechanics (MM), or empirical force field methods (EFF), are so called because they are a model based on equations from Newtonian mechanics. This model assumes that atoms are hard spheres attached by networks of springs, with discrete force constants. The force constants in the equations are adjusted empirically to reproduce experimental observations. The net result is a model which relates the "mechanical" forces within a structure to its properties. Force fields are made up of sets of equations each of which represents an element of the decomposition of the total energy of a system (not a quantum mechanical energy, but a classical mechanical one). The sum of the components is called the force field energy, or steric energy, which also routinely includes the electrostatic energy components. Typically, the steric energy is expressed as

$$E_{\text{total}} = E_{\text{steric}} + E_{\text{electrostatic}} + E_{\text{bonds}} + E_{\text{angles}} + E_{\text{vdW}} + E_{\text{torsion}} + E_{\text{charge-dipole}}$$

The overall form of each of these equations is fairly simple, ie, energy = a constant times a displacement. In most cases the focus is on differences in energy, because these are the quantities which help discriminate reactivity among similar structures. The computational requirement for molecular mechanics calculations grows as n^2 , where n is the number of atoms, not the number of electrons or basis functions. Immediately it can be seen that these calculations will be much faster than an equivalent quantum mechanical study. The size of the systems which can be studied can also substantially eclipse those studied by quantum mechanics.

In a force field calculation, a molecule in three dimensions is constructed using either Cartesian coordinates x , y , and z or via an internal coordinate matrix consisting of bond distances, bond angles, and dihedral angles to specify the atoms' unique positions. Then the initial structure is evaluated to determine the extent to which each degree of freedom (bonds, angles, etc) deviates from the ideal (the zero-energy value) for the particular element and its hybridization. An energy minimization process follows wherein the energy associated with the distortions from ideal is minimized as the individual atomic positions or degrees of freedom are adjusted. Iteratively, this converges on a "minimum energy" or an "optimized" structure. This structure represents the best attempt of the minimization algorithm to render the smallest deviations in position of each of the atoms such that either the derivatives of the change in energy associated with the deviations are the smallest, or they satisfy either energetic convergence or coordinate change criteria from iteration to iteration. It should be noted that this process is analogous to the geometry optimization process within a quantum mechanical program, except that there the objective is to converge on a structure which yields the smallest energy derivatives and lowest total energy from solution of the SCF equations. Most simple molecular mechanics force fields include terms (Fig. 8) for:

bond stretching: $E_l = k_l(l_0 - l)$

bond angle distortion (bending): $E_\theta = k_\theta(\theta_0 - \theta)$

dihedral angles: $E_\tau = V_n/2(1 + \cos n\omega)$

van der Waals nonbonded interactions (1 ... 4 interactions or greater): $E_{\text{vdW}} = C/r_k^6 - A/r_k^{12}$,

Coulombic interactions where q_k can be dipoles or charges $E_{\text{Coul}} = q_k q_l / r_{kl}$.

In these equations, k_l , k_θ , V_n , C , and A represent the empirically adjusted constants associated with changes in bond lengths, angles, dihedrals, and nonbonded interactions. These terms plus 1 ... 3 nonbonded interactions, and cross-terms such as stretch-bend are considered in the most complex computational models as well as in the experimental force fields. Practically all of the early molecular mechanics force fields utilized force constants directly from vibrational spectroscopic studies (see background in Refs. 59 and 60). That is, for a particular interaction included in the force field, the force constant applied to an interaction was one which had been experimentally determined. Although this method can be utilized, it is very difficult to develop a generalized force field for a broad spectrum of molecules, because not all experimental force fields are derived to the same level of accuracy, nor are they consistent, ie, having all force constants derived concurrently. For this reason Allinger (60) has suggested:

... we must not look at a force field calculation and ask "what interactions are really occurring in the molecule"; rather the question must be "what interactions are really occurring in *our model* of the molecule." Of course the hope is that the answers to the last question will, in fact, converge upon the answers to the first question as force fields improve.

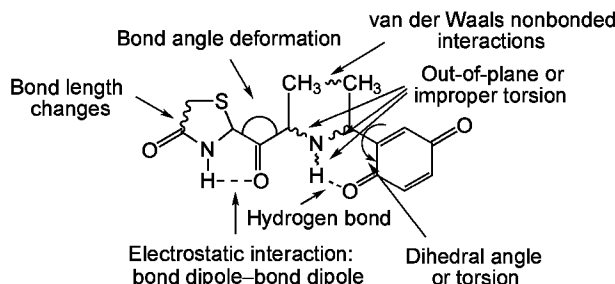


Fig. 8. Structural representation of the energetic components of a typical molecular mechanics force field.

A more detailed account of the MMI force field is also available (61).

Force Fields, Molecular Dynamics, and Vibrational Spectroscopy. The details of the relationship between molecular mechanics force fields and spectroscopic vibrational force fields have been discussed (60). Other fundamental papers on molecular mechanics are available as well (59,62–63,65–67). The link between molecular mechanics and molecular dynamics comes about through the force field itself. In molecular mechanics, the main interest here is in computing the energy of molecules in the gas phase at room temperature in a single, discrete configuration and conformation; time is not a variable in the equations. The goal is to know the structure of the molecule, bond lengths, angles, etc; conformational energy relative to some reference structure (eg, *gauche* vs *anti*); the heat of formation; and, perhaps, the vibrational frequencies. From molecular dynamics, the objectives are properties which represent a time-averaged ensemble of states, including, for example, conformationally excited states, rotational states, interconversion rates, or inversion barriers for amines or amides. From this ensemble of states, characterizing the existence of the excited states and their contribution to the total energy of the system (from a Boltzmann distribution) is next. From an understanding of the vibrational spectroscopic roots (64) of molecular mechanical force fields used in dynamics simulations it can easily be appreciated that molecular dynamics may be seen as an extension bridge between theory and experiment, linking "static" molecular mechanical representations of properties with "dynamical" experimental properties.

The accuracy of molecular mechanics and that of molecular dynamics simulations share an inexorable dependence on the rigor of the force field used to elaborate the properties of interest. This aspect of molecular modeling can easily fill a volume by itself. The topic of force field development, or force field parameterization, although primarily a mathematical fitting process, represents a rigorous and highly subjective aspect of the discipline (68). A perspective behind this high degree of rigor has been summarized (69). Briefly put, the different schools of thought regarding the development of force fields arose principally from the initial objectives of the developers. For example, in the late 1960s through the 1970s, the Allinger school targeted the computation of the structure and energetics of small organic and medicinal compounds (68,70,71). These efforts involved an incremental development of the force field, building up from hydrocarbons and adding new functional groups after certain performance criteria were met, eg, reproduction of experimental structures, conformational energies, rotational barriers, and heats of formation. Unlike the consistent force field approach of Lifson and co-workers (59,62–63,65), the early Allinger force fields treated a dozen or more functional groups simultaneously, and were not derived by an analytical least squares fit to all the data (61). However, because the focus of Lifson was the analysis and prediction of the properties of hydrocarbons or peptides, it was not surprising that a consistent force field was possible. The number of variables to be optimized concurrently to permit calculation of all the structure elements, conformational energies, and vibrational spectra concurrently was, and still is, a massive quantity. However, the calculation for a limited number of functional groups could be accomplished, albeit slowly. If the goal is to reproduce and predict vibrational spectra, the full second derivative force

constant matrix must be computed. Doing so in the early 1970s meant that a researcher was limited severely by the computing resources then available. By omitting or significantly reducing the number of the second derivatives computed, Allinger and co-workers were able to address a much larger number of problems in conformational analysis while producing force fields which were robust enough to be useful many years after their introduction; information regarding the utility of MM2 is available (72). What is also not surprising is that many of the schools of force field development have converged significantly. Today (ca 1997), the most extensively developed force fields rely heavily on *ab initio* quantum mechanical calculations for both structural and energetic data not available from experiment. Examples include the force fields from the Hagler group (73) (CFF9X CVFF, Biosym MSI), the Halgren MMFF from Merck (74), and Allinger's MM4 (75). Unlike what was the case in 1970, today (ca 1997) the volume of experimental data is vastly dwarfed by the requirements placed on force fields for both molecular mechanics and molecular dynamics simulations. Researchers have endeavored to satisfy the requirements of both breadth of application and depth of rigor for each structural class (73–75). However, there remain families of force fields which can be classified as special-purpose force fields; that is, the developers chose to reproduce and predict the properties of specific structural classes, such as peptides and proteins (73,76,77), nucleic acids (76), or organometallic compounds (78). The utility of these force fields should not be underestimated, because they are usually developed by researchers who have significant experience in their respective fields and are designed to serve the special or idiosyncratic requirements of a subspecialty of a particular discipline. Because these techniques are specialized in their focus, care should be exercised in using such methods beyond their intended scope. As has been relevantly noted before, "Models are to be used, not believed" (79).

This leads to the questions as to how well a given force field performs and whether a force field is appropriate for the problem at hand. With regard to the former, a number of perspectives have been given (61,80,81). A series of publications has detailed how to generate highly accurate force fields, including corrections and extensions to an earlier force field (81–82). These studies provided a basis for additional force field development and rigorous application of the MM2 method to the prediction and understanding of chemical reactions (83). Quantitative assessments (84) provide the researcher with data sufficient to evaluate whether the methods in question are suitably accurate for the specific needs. Clearly, if only a reasonable 3-D structure is needed, to give a spatial perspective, then highly accurate bond lengths and angles are not necessary. However, if a force field is being used to assist in the refinement of an experimental structure by nmr or x-ray methods, then it is desirable to use the force field which provides the most accurate assessment of structure and energies for the class(es) of structure under evaluation. Boyd (85), in several articles in the series he has co-edited, has discussed extensively the types of molecular mechanics programs available and included background information to help guide the novice and expert alike in the choice of the appropriate computational tools to solve individual problems. Applications of force field techniques to problems in environmental chemistry, materials science, and molecular biology demand that the methods go substantially beyond those for which reliable experimental data is available. Only since the time computers have made larger-scale *ab initio* quantum mechanical calculations practical for appropriate model systems have these techniques been reliable for such a broad spectrum of important applications.

Molecular Dynamics and Monte Carlo Simulations. At the heart of the method of molecular dynamics is a simulation model consisting of potential energy functions, or force fields. Molecular dynamics calculations represent a deterministic method, ie, one based on the assumption that atoms move according to laws of Newtonian mechanics. Molecular dynamics simulations can be performed for short time-periods, eg, 50–100 picoseconds, to examine localized very high frequency motions, such as bond length distortions, or, over much longer periods of time, eg, 500–2000 ps, in order to derive equilibrium properties. It is worthwhile to summarize what properties researchers can expect to evaluate by performing molecular simulations:

- Conformational states and energetics
- Kinetic properties: rates of reaction and interconversion
- Reaction pathways
- Solubilities
- Diffusion rates
- Binding and complexation data
- Folding processes
- Transition temperatures
- Free energies for point mutations
- Free energies of binding
- As an adjunct to x-ray and nmr for structure refinements

As noted, force fields are a set of equations relating the total energy of the system to its individual interaction components. Popular force fields for simulations on organic and biomolecules include the following:

Program	Principal author(s)
AMBER	Kollman Weiner Singh Pearlman
CHARMM CHARMM	Karplus Brooks MSI
CFF9X	Hagler BIOSYM
GROMOS	van Gunsteren
MM2 MM3	Allinger Welsh
BOSS (Monte Carlo)	Jorgensen Tirado-Rives

After selection of a force field simulation program which is appropriate to a given problem, the general procedure is as follows:

- (1) Initial structure equilibration, wherein bad or close contacts are relieved; this may be done with constraints on bonds, eg, to simplify the process (the premise of the SHAKE technique).
- (2) Structure refinement (locating energy minima) by energy minimization: this step may take a few hundred picoseconds of simulation time if the original configuration is far from a minimum. Note that the researcher may converge to a local minimum which is significantly higher in energy than the global minimum.
- (3) Techniques used to find global and local energy minima include: sequential simplex, steepest descents, conjugate gradient and variants (BFGS), and the Newton and modified Newton methods (Newton-Raphson).
- (4) Set up and solve Newton's equation, equation 5, for each atom in the system:

$$F = m_i a_i(t) = m_i \partial^2 r_i(t) / \partial t^2$$

(5)

where F = force on atom i at time t , a = acceleration of atom i at time t , and r = position of the atom i at time t .

- (5) Evaluate the force F as the negative gradient of the potential energy function:

$$F_i = - \partial v(r_1, r_2, \dots, r_N) / \partial r_i$$

(6)

- (6) Compute normal modes. These represent primarily hammonic motions internal to the molecule. There are $3N-6$ displacement eigenvectors, where N is the number of degrees of freedom of the system. The associated eigenvalues are the frequencies.

A variety of techniques have been detailed for handling Newton's equations of motion, equation 5 (86–90). Integration techniques yield atomic

positional velocities which are used graphically to display the internal motions and paths of motion computed during the simulation runs. Understanding the behavior of a molecular system as a function of time is a critical element in any simulation, and velocities can be collected into a series snapshots called a trajectory (atomic coordinate sets). From trajectories, researchers can determine the level of cooperation of motions in the folding or conformational processes which a polymer or biopolymer chain might undergo, eg, the interconversion of forms of DNA prior to and during complexation with a protein. Although such simulations represent a full-scale challenge to the simulation technique owing to the wide variety of atom types and associated parameters, a different but no less rigorous type of challenge exists in the simulations of simple polymers of polyethylene. For example, in studies on polyethylene chains of size C_{10} to C_{100} (91), dynamics were used to elucidate the cooperation of motions at neighboring rotational sites, with a finding that for very short time periods the torsional motions of the chains are effectively Brownian in behavior, as had been found earlier. It should be noted that the operational methodology for applying molecular dynamics does vary from application to application, but it is probably reasonable to assume that simulations involving most polymeric systems other than peptides can be addressed similarly, with appropriate modifications for atomic species and periodicity (92). Even peptide systems are addressable by similar methods, despite the breadth of the literature on it that is available suggesting otherwise. The proliferation of dynamics programs for handling peptides and proteins may be the result more of differing perspectives on the method of sampling of excited states, energy refinement algorithms, implementation of constraints, or restraints during the simulation, or force field. The situation is similar to understanding that a car is a vehicle for transporting people and cargo from point A to point B. Whereas the relative comfort and efficiency of the process may vary with the specific vehicle manufacturer, the fact that cars transport people and cargo is common to all manufacturers. All of the subsequent levels of structure and function in peptides and proteins depend intrinsically on the primary structure of the system. Unlike simulations on homopolymers or rigid systems with regular periodic sets of atoms at essentially fixed distances, as in the case of zeolites, the complexity of the simulation problem increases as a function of the variety of the types of structural motifs (helices, sheets, turns, barrels, coils) present, because each motif can behave both as individual atomic entities and as a single collective entity. Numerous investigators have detailed approaches to dealing with the analysis of these systems (73,88–90,93,94). Dynamical characterization of biopolymers and related systems (ca 1997) remains one of the most challenging and stimulating aspects of molecular modeling.

- In the context of molecular dynamics, some additional pointers may help give a fuller perspective on the MD simulation operational process:
- Initializing the initial kinetic energy and temperature of the system: it is necessary to start the motion at some level, eg, assume a Boltzmann (random) distribution of atomic velocities, at 300 K.
- Timesteps for the simulation need to be short enough to capture high frequency motions such as bond stretching, eg, 10^{-15} s.
- Simulations need a substantial number of timesteps to sample configurational states such that desired properties are represented well enough to either converge or not –100 ps. For systems of >100 atoms, run for 500–2,000 ps if computing resources permit.
- Langevin dynamics: a technique to reduce the total number of equations of motion that are solved. Utilize the Coupled Heat Bath, wherein the method models the solvent effect by incorporating a friction constant into the overall expression for the force.
- The SHAKE method for bond constraints reduces the number of degrees of freedom during the initial stages of simulations; it is good for minimizing solvent bath overhead.
- Simulated annealing is a technique which heats up the system and slowly cools it, perhaps to multiple different minima.
- Use nonbonded (NB) truncation methods to reduce size of NB pairlist; it is a dominating term in the calculation. It is important to remember the pairlist is N^2 , consider truncation of NBs at 100–120 nm (10–12E), and to experiment with electrostatic cutoffs independently of van der Waals.
- Update the NB list every 25–50 timesteps.
- Utilize periodic boundary conditions, which permit reduction of the number of nonbonded interactions at greater distances by involving only the "nearest neighbor" atoms from copies of the system which are in different but adjacent cells.
- Use a solvent or water bath to capture the influence of solvent on the solute. Use the fewest shells of water solvent possible, but no fewer than two, if resources are scarce. Use a formal "box" of water, if possible, to reduce the influence of edge effects.

Monte Carlo (MC) techniques for molecular simulations have a long and rich history, and have been used to a great extent in studying the chemical physics of polymers. The majority of molecular modeling studies today do not involve the use of MC methods; however, the sampling capability provided by MC methods has gained some popularity among computational chemists as a result of various studies (95–97). Relevant concepts of MC are summarized herein.

Monte Carlo methods as applied to chemical problems owe their popularity to the work of Metropolis (98), who first utilized them on very early computers to evaluate properties of simple molecular systems. Typically, a set of configurations in a given thermodynamic ensemble is generated by the random sampling of configuration space. By configurations is meant sets of atomic coordinates corresponding to discrete geometries, including those with substantial distortions in bonding or other periodic structural behaviors. In MC calculations, no time relationship exists between successively calculated configurations. Unlike in a molecular dynamics simulation, there is no path or trajectory which the system follows as interconversions occur and states are sampled. Rather, in MC simulations random steps are taken, necessitating that a very large sampling be done to optimize the sampling of the desired configurations. Unfortunately, with the complexity of the potential energy surface of large polymers or proteins, it is not always possible to be assured that one has sampled sufficiently. This open-endedness may be the primary reason that molecular modelers have not embraced Monte Carlo techniques as freely as they have molecular dynamics (MD) simulations, despite the steep computing requirements of the latter.

For many systems the ensemble that is used in an MC simulation refers to the canonical ensemble, (N, V, T) . This ensemble permits a rise and fall in the pressure of the system, P , because the temperature and volume are held constant. Thus, the probability that any system of N particles, in a volume V at temperature T is found in a configuration x is proportional to the Boltzmann weighted energy at that state, E_x , and it is given by

$$P_x = \frac{\exp(-E_x/kT)}{Z} \quad (N!Z)$$

where Z is the configurational partition function

$$Z = \frac{1}{(N!)^2} \int \exp(-E_x/kT) dx = \exp(-A/kT)$$

and A is the Helmholtz free energy. The average value of a property is

$$\langle F \rangle = \int F_x P_x dx$$

where F_x is the property evaluated at configuration x . Another popular ensemble used in protein and DNA MC simulations (95) is the (N, P, T) ensemble.

One method used to enhance the efficiency of sampling is biased sampling (98). An algorithm utilizing biased sampling allows low energy configurations to be sampled more often, and is usually more efficient than random sampling at sampling those configurations which contribute more significantly to the true average of a system. For example, to simulate an MC run (92), an algorithm might involve the following steps: (1) calculate the energy of the current configuration, E_1 ; (2) assign a new configuration, with a new energy, E_2 ; (3) calculate the weight, Ω , of $\exp(-[E_2 - E_1]/kT)$; (4) compare the result to a uniform, random value in the range of [0,1]; and (5) if the weight Ω is greater than the random value, reset E_1 to E_2 , and restart the cycle. The Markov chain that results from this process produces a probability of a given configuration that is proportional to the calculated weight.

The list which follows gives an outline of the properties of a Monte Carlo simulation used in the context of molecular modeling studies for sampling either multiple conformations of smaller, flexible structures or multiple local minima of larger macromolecules or polymers:

- Probabilistic, as opposed to deterministic molecular dynamics

Essentially random atomic moves use the Metropolis method (98)
 Calculate new energy and compare to previous configuration(s)
 Keep or discard new configuration
 Develop a statistical ensemble of energetically accessible states
 Usually sample many millions of states; but not always easy to sample space well
 Use Boltzmann or biased-sampling techniques
 Umbrella sampling can give free-energy differences, but not absolute free energies
 Usually done in NPT-isothermal isobaric ensembles, including a water box
 Useful for efficient sampling of conformational space in systems such as polymers (92) and peptides (93–95).

Combined Quantum and Molecular Mechanical Simulations. A recently developed technique is one wherein a molecular dynamics simulation includes the treatment of some part of the system with a quantum mechanical technique. This approach, QM/MM, is similar to the coupled quantum and molecular mechanical methods introduced by Warshel and Karplus (45) and at the heart of the MMI, MMP2, and MM3 programs by Allinger (60). These latter programs use quantum mechanical methods to treat the π -systems of the structures in question separately from the sigma framework. The results are combined at the end to render a structure which is optimized and energy-refined to satisfy both SCF and force field energy convergence. Newer QM/MM approaches treat the bulk of the molecule via a molecular mechanical force field to give its dynamical behavior, whereas a selected portion of the structure(s) is treated quantum mechanically to yield information about interactions between the selected segments. This technique is particularly appropriate where applied to systems such as ligands bound to biological receptors or molecular "guests" trapped by "hosts."

In the 1980s the most comprehensive implementation of this method was the program Gaussian80 (UCSF) developed at the University of California at San Francisco (99), which combined the molecular mechanics and dynamics code AMBER (100) and the *ab initio* quantum mechanics program Gaussian80 (101). The program permits researchers to generate structures of interest, then run initial steps of energy minimization to resolve close contacts and significantly distorted bond lengths and angles. The quantum mechanical atoms of the enzyme-substrate complex would be defined next. The dynamics simulation would be run over several hundred picoseconds, with the majority of the atoms dealt with using the MM component of the program; substructures including the quantum mechanical atoms are treated by a full-scale *ab initio* calculation at the STO-3G level, eg, at selected intervals. The process would be captured in a trajectory file, and the quantum mechanical data would permit researchers to investigate reaction mechanisms, proton shuttling, salt-bridge formation, and other intermolecular associations pertinent to the dynamical processes. The general method has been elaborated by a number of researchers, a published survey (102) provides further details and examples of the application of the method.

Treatment of QM/MM calculations would not be complete without mention of the use of a newer quantum mechanical method called density functional techniques. Although density functional theory (DFT) is not new, recent implementations have made it significantly more popular than in the past. In particular, the potential of DFT as the QM component of combined QM/MM efforts has focused more attention of DFT methods; an introduction to DFT and its applications is available (103) which cites the primary literature in DFT, including a classic monograph (104). The comprehensive survey of QM/MM (102) also gives references pertinent to the application of the DFT method to QM/MM studies. Briefly put, DFT methods depend on the use of a functional of the electron density rather than the wavefunction functional, the advantage being that the density is an observable entity whereas the wavefunction is not. Further simplifying the calculations in DFT is the fact that the electron density has three spatial coordinates, regardless of the number of electrons in the chemical system. This makes it possible readily to compute properties directly from the electron density for systems of hundreds of electrons, a feat not easily accomplished routinely within wavefunctional theory. Just as occurs in the case of *ab initio* Hartree-Fock theory, DFT utilizes basis sets, and many of the more popular ones, along with the variants of DFT functionals, are available in commercial QM programs such as the Gaussian series (101) or SPARTAN (105).

Structure Generation, QSAR, and CoMFA Modeling Methods.

Rule-Based Structure Generation Techniques. Figure 1 summarized the process for small molecule modeling. An important component of that process is the generation of the three-dimensional coordinates of the molecules to be studied. Many modeling program suites have their own 2-D drawing-to-3-D-coordinates conversion routines. However, it was not always so simple or routine to make the 2-D-to-3-D conversions. During the 1980s, significant progress toward simplifying this aspect of modeling was made, particularly in the area of knowledge-based methods. One of the most rigorous, and by far the most successful, of the conversion programs is CONCORD (106), produced at the University of Texas at Austin. CONCORD is a rule-based algorithm which evaluates atomic connectivity tables and maps reasonable bond lengths, angles, and dihedrals to the structures based on connectivity information and other geometric features permissible for a given atom and its normal valence states. The tool has become the de facto standard for the conversion of databases of hundreds to millions of 2-D structure representations into 3-D structures. Because the geometric parameters for structures were derived from AM1 (Austin Model 1, a successor to MINDO) semiempirical quantum mechanical optimized geometries for families of model functionalities, the method is quite extensible and reliable, although it does suffer from some of the same deficiencies which are characteristic of AM1 calculations (eg, underestimated dipole moments, incorrect conformations for hydroxyl groups and bioxyl groups). A thorough review of methods for generating 3-D structures is available (107).

Distance Geometry. Another technique which utilizes both rule-based and computational geometry methods for generating 3-D structures is called distance geometry. This technique, which owes its origins to Cayley (108) in 1841, has been popularized as a result of pioneering work (109,110) and by the DGEOM program (111), and applications (112,113) developed by researchers at the University of California at San Francisco. The technique is unique among structure or conformation generation methods in that it applies equally well to both small and large molecular ensembles. The distance geometry algorithm has been interfaced with a graphical display engine (113) and the method used to generate multiple conformations of flexible and complex ring systems to facilitate conformational analysis. The use of distance geometry (DG) in conformational searching has been eloquently summarized (114). There, the idea is to generate multiple, geometrically feasible conformations for energy refinement by MM or QM methods. The mix of random and systematic structure generation permits a fairly thorough search of conformational space. Others have also detailed the use of DG in molecular modeling, particularly in the context of macromolecular structure generation and docking (115), wherein the basics of the mathematical elements of the technique, how distance constraints are developed for determining the distance bounds which determine structural features, the metric matrix technique for developing the geometries, and additional techniques from computational geometry which are integral to DG are all given. Of particular interest are the summaries of the applications of DG to the generation of polymers and biomolecules (alone or in conjunction with nmr or x-ray) data, ligand-receptor docking studies, and pharmacophore modeling. Ghose and Crippen (116) have provided a summary of the spectrum of distance geometry applications and the utility of distance geometry in molecular modeling, a portion of which is quoted in a slightly adapted form in the following list:

- (i) Generate an approximate three-dimensional structure of a ligand.
- (ii) Generate the low energy conformations ...
- (iii) Evaluate the minimum and maximum distances between the ligand atoms in [the structure's] energetically allowed conformations.
- (iv) [Using a computerized search method, develop] a hypothesis [of] the binding mode of the ligand at the receptor site ...
- (vii) Evaluate the alternat[ive ligand] binding modes due to rigid rotations ...
- (viii) Classify the site pockets into different types to differentiate [the level of] interaction [with ligands]
- ...

In addition, researchers can survey the surface and clefts of a macromolecule for potential receptor sites based on ligand distance ranges; "use the common distance range of the superimposed atoms" (116) to facilitate the development of a "pharmacophore" or critical contact assembly; besides the ligand

binding modes due to rigid rotations (116), evaluate those due to conformational flexibility, ie, by using highly constrained ligands and only considering the lowest energy conformations of flexible ones, the scope of this process can be reduced; evaluate the binding data with a training set and develop a regression model to predict the binding of other known ligands; and finally, "if the predictive power is acceptable, use it to predict [or screen sets of] structure" (116) from databases or combinatorial libraries.

Quantitative Structure–Activity Relationships (QSAR). Quantitative Structure–Activity Relationships (QSAR) is the name given to a broad spectrum of modeling methods which attempt to relate the biological activities of molecules to specific structural features, and do so in a quantitative manner (see ENZYME INHIBITORS). The method has been extensively applied. The concepts involved in QSAR studies and a brief overview of the methodology and applications are given here.

Historically, QSAR has been applied primarily to drug molecules; however, more recently, Quantitative Structure–Property (Toxicology) Relationships (QSP (T) R) have been elaborated by a number of researchers (117). Regardless of the context of application, QSAR has its roots in the use of Linear Free Energy Relationship (LFER) methods of deriving an equation which correlates the observed effect and the structural features responsible for the activity. The work of Hansch in relating hydrophobicity to biological activity is exemplary as a pioneering application of LFER in QSAR (118). It is interesting to note, however, that many of the early investigations in QSAR involved analysis of the relationship of hydrophobicity to activity, the nature of which relationship is more often parabolic rather than linear (119,120). QSARs are usually best derived from a series of compounds (typically differing only at one or two substitution points) for which the activities are well-determined by a stable biological assay. A QSAR table can be established wherein the columns are assigned to activities (the *ys*), and the metrics or properties (the *xs*), which can be either observable properties such as log *P*, hplc retention times; nmr chemical shifts; computed values such as shape and size descriptors; dipole moments; atomic charges; or conformational energies. Each row represents an individual compound or conformation. Statistical relationships can then be developed from this table by means of univariate or multivariate techniques such as linear or multiple linear regression (MLR), or partial least squares (PLS). If the statistical significance of the relationship is sufficiently high, then this relationship is robust enough to be used to predict or assess the activity of untested compounds. Usually the known data is divided into two groups, a training set and a test set to establish the statistical model. As a rule of thumb, there should be between 3 and 10 times the number of observations (rows) as *x*-variables in order to derive a model which would have predictive power and be able to minimize chance correlations. Under these circumstances, there would be some constraints against pursuing QSARs if only a few observations (molecules for which activities or properties are known) are available. Often this is the case, yet investigators have proceeded with the development of the QSAR. A discussion of the use of Partial Least Squares (PLS) as a potential means of resolution for this difficult is treated herein in the context of the CoMFA paradigm. It should also be recognized that many more molecular conformations and property descriptors than ever before can now be computed. For example, in the QSAR and Diversity modules of the Cerius2 software from Molecular Simulations, Inc. (121), there are approximately 160 shape, size, and electrostatics descriptors which can be computed. This relative abundance is in sharp contrast to the small numbers of descriptors available to early investigators, who worked diligently with classical Hammett σ s and log *P* values to derive LFERs (122–125). A particularly lucid account of the Hansch Approach to application of LFERs in QSAR which is illustrated with numerous examples is given by the Hansch collaborator and specialist in QSAR, T. Fujita (126). An example of the classical linear equation is represented by equation 7:

$$\text{PED}_{50} = 0.606(\pm 0.184) E_s^c + 1.518(\pm 0.265) \tag{7}$$

where *n* = 16, *r* = 0.884, and *s* = 0.204. This equation was derived for a series of fungicide compounds of the type *N*-substituted aminoacetonitriles, RNHCH₂CN, wherein it was found that the potency (ED₅₀ = Effective Dose to kill 50%) was dependent on the corrected steric parameter, *E*_{*s*}^{*c*} (127). This variant of the Taft steric parameter, *E*_{*s*}, emphasizes the effect of branching in addition to steric bulk of the R group. Another example of a simple linear equation has been derived for the enzymatic hydrolysis of esters by the serine hydrolase, trypsin, (128). For a series of esters, X–Ph–OCOCH₂NHCOPh, hydrolysis yields equation 8, where X is 4-SO₂NH₂, 4-NH₂, 4-CN, 4-NO₂, 4-NHCONH₂, 4-OCH₃, H, 4-CH₃, 4-Cl:

$$\log(1 - K_m) = 0.71(\pm 0.17) \sigma + 3.31(\pm 0.09) \tag{8}$$

where *n* = 10, *r* = 0.961, and *s* = 0.100. These examples are not atypical of the hundreds which can be found in the literature. More exemplary of the relationships involving the log *P* are given by equation 9:

$$\log(1 - C) = a(\log P) + b(\log P)^2 + c \tag{9}$$

where *C* is the equipotent concentration or dose, and *a*, *b*, and *c* are the coefficients of the linear, quadratic, and constants terms, respectively. Often, the linear term is zero, and a purely parabolic relationship between activity and log *P* is observed (126). More complex multivariate linear, bilinear, and parabolic equations can be valid for QSARs. However, it is important to perform a critical evaluation of both the biological data and the statistical paradigms before publishing such results. The treatment and testing of both the biological data and the QSAR model equations have been detailed (129–132). More recent studies on this subject can be found in the chapter by Pleiss and Unger (133), and in the journal, *Quantitative Structure–Activity Relationships* (Wiley-VCH).

The range of application of QSAR is extensive. Additional aspects of QSAR which are important and have recently reached the production level of application in industrial QSAR studies (134) include molecular similarity searching, 2-D fingerprinting, 3-D substructure searching, molecular superpositioning, pharmacophore identification, pharmacophore searching, 3-D databases, structural alignments, receptor–ligand binding energetics modes, 3-D QSAR, molecular shape analysis, 1-D, 2-D and 3-D descriptors, including shape size, charge, and hydrophobic fields, *de novo* ligand design, and hypothetical active site lattices.

The entire domain of "new-lead" discovery has expanded considerably. This development has affected what have traditionally been divergent approaches, namely QSAR and structure-based design, leading them to become integrated so as to provide a more powerful approach (135). Several recently published comprehensive volumes capture the state of the art and can be consulted to determine precedents relevant to any particular study (134–137).

Comparative Molecular Field Analysis (CoMFA). Another method for molecular modeling is one which was developed in the mid-1980s but has come into frequent and practical use in the 1990s owing to the rapid advances in workstation computing power and in techniques for aligning diverse sets of structures. This technique, developed by Cramer (138) and frequently referred to as 3-D QSAR, is called Comparative Molecular Field Analysis (CoMFA) (see ENZYME INHIBITORS). CoMFA is a computational technique which attempts to mimic the interaction of a "ligand" with a "receptor" by means of a lattice of points (receptor) within which the molecule of interest (ligand) is placed, and the interactions between the molecule and the grid points are evaluated. The points of the lattice which are inclusive of the volume of the molecule of interest are discarded. The researcher must propose an alignment (superpositioning) of the structures of interest. This can be done by rigid or flexible fitting, in either case an additional QSAR descriptor, the distortion energy of superpositioning, can be added to the classical QSAR regression. All other points are assigned both steric and electrostatic properties, such as *C*_{*sp*}² carbon van der Waals steric properties, and a +1 charge for electrostatics. The interaction energies for both property types are then evaluated for the approximately 2,000 points which remain in the lattice. These data (the *xs*) are then processed by Wold's technique of partial least squares (PLS) regression analysis (139) against the activity values (the *ys*) supplied by the researcher. The program produces a model which will both reproduce the training set of data values (bioactivity, or a similar property) and have predictive power as well. The PLS model is cross-validated by successively eliminating observations, rederiving the model, and predicting the eliminated observations. As new, diverse data are added to the training set, the predictive power of the model is enhanced. Additionally, the CoMFA model is visually displayed, indicating regions either where steric bulk is favored or less so, as well as where changes in the electrostatic field enhance or diminish activity. The power of CoMFA is made possible by the PLS technique, because without PLS to reduce the dimensionality of the *xs*, it would not be possible to derive a believable regression model when presented with a data matrix which is 10–20 rows by 2,000 columns wide. Studies on factor analysis and principal components regression (140), on chemometric tools (141), on

the performance of biased regression techniques (142) have helped in understanding the complexity of regression techniques, their pitfalls, and their importance of QSAR. Indeed, a set of recommendations and caveats regarding the use of CoMFA has been published (143), and as with any computational or modeling technique, its capabilities and limitations should become better understood as the frequency of its use grows. In summary, QSAR techniques give researchers a wide range of approaches to the problem of quantitating the relationship between chemical structures, changes in chemical structures, and observed physical and biological properties.

Summary

Clearly, whereas molecular modeling as a practice has its roots in the development of quantum theory at the turn of the twentieth century, it has been the exponential growth in computing power between the mid-1970s and the mid-1990s that has catalyzed the development and application of molecular modeling methods during that period. The spectrum of software systems available (144) covers all aspects of modeling. A sampling is given in Table 1. Although it is not possible to duplicate the exhaustive survey of available systems given in Reference 144, Table 1 furnishes a starting point from which to search for the right tool to address specific problems.

Table 1. Molecular Modeling Software Systems^a

System(s) acronym(s)	System name or type	System producer
<i>Databases and DB management systems</i>		
MACCS CCDC	Cambridge Crystallo-graphic Data Center X-ray Databases	Daylight Chemical Infor-mation Systems Molecular Design Limited, Inc. (MDLI) Cambridge Crystallo-graphic Data Center
PDB CAST-3D UNITY	Brookhaven Protein Data Bank The Chemical Abstracts Service 3-D Structure DB Module of the SYBYL suite	Brookhaven National Laboratory Chemical Abstracts Service Tripos
<i>Desktop modeling and data management</i>		
ISIS Draw, SAR, Excel AccuModel SAS, JMP, MINITAB ChemDraw, Chem3-D, ChemOffice Alchemy 2000 CaChe	ISIS Add-ins statistics	Molecular Design Limited, Inc. Microsimulations, Inc. SAS CambridgeSoft Tripos Oxford Molecular
<i>Standalone software programs (workstation versions)</i>		
NBO, GAUSSIAN, SPARTAN, MOPAC, PROAIM PEOE Methods for Charges MM3, MM4, MMFF AMBER, GROMOS, QUANTA CHARMm BOSS	quantum mechanical molecular mechanical molecular dynamics Monte Carlo simulations	Gasteiger, Scheraga, Gombar (HDI) Allinger Halgren (see text) Jorgensen
<i>Molecular modeling software suites (workstation)</i>		
Chem-X Hyperchem InsightII, Cerius MacroModel OMG SYBYL		Chemical Design, Ltd. Hypercube, Inc. (Ostlund and co-workers) Molecular Simulations Biosym Columbia University (W. C. Still) Oxford Molecular Tripos
<i>2-D to 3-D converters</i>		
CONCORD CORINA WIZARD DGEOM	converter conformation generator converter conformation generator converter conformation generator distance geometry program <i>QSAR</i>	Tripos Gasteiger OMG Dolata & Leach QCPE (UCSF) DuPont)
ADAPT APEX CHEMEST CLOGP MOLCONN-X QSAR-PC SigmaStat TOPKAT TOPMOST	pattern recognition toolkit full range of statistical treatments for QSAR Kier & Hall Indices stat analysis package toxicology prediction tool computes charges, descrip-tors for QS(APT)R <i>Other applications</i>	P.C. Jurs (MDLI) MSI Technical Database Services Biobyte, Inc. (Pomona MedChem Project) L. Hall Biosoft, Inc. Jandel Scientific Software Health Designs Inc. (HDI) HDI
AutoDOCK CAVEAT DISCO	Monte Carlo docking of ligands to receptors database generator for new ligands tool for deriving pharma-cophore from active compounds	A. Olson at Scripps Institute P. Bartlett at University of California, Berkeley Y. Martin at Abbott Labs; available from Tripos
DOCK	ligand docking active site probe tool	I. D. Kuntz at University of California, San Fran-cisco
GRIN GRID HINT LEAPFROG LUDI	nonbonded force probe of active sites adjunct to CoMFA, computes hydrophobic fields generates new leads from fragments <i>de novo</i> ligand design	Molecular Discovery, Ltd. (Goodford) EduSoft, Inc. (Kellog) Abraham) Tripos MSI (Brñhm)

^a Ref. 144.

An understanding of the precedents in both methods development and applications citations in the literature is thus critical to the researcher working in fields that employ molecular modeling as a tool. With it, the varied application and untapped potential of molecular modeling may be used more profitably in individual researchers' specific fields of interest.

BIBLIOGRAPHY

1. W. Heisenberg, *Z. Phys.* **33**, 879 (1925).
2. E. Schrödinger, *Ann. Phys.* **79**, 361, 489; **80**, 437; **81**, 109 (1926).
3. J. D. Bolcer and R. B. Hermann, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. 5, VCH Publishers, New York, 1994, pp. 1–63.
4. R. S. Mulliken, in D. A. Ramsey and J. Hinze, eds., *Selected Papers of Robert Mulliken*, University of Chicago Press, Chicago, 1975, pp. 39–42.
5. R. S. Mulliken, *Life of a Scientist*, Springer-Verlag, Berlin, 1989, pp. 136–161.
6. F. London, *Z. Phys.* **63**, 245 (1930).
7. E. Hückel, *Z. Phys.* **70**, 204 (1931).
8. J. D. Kemp and K. S. Pitzer, *J. Chem. Phys.* **14**, 733 (1936).
9. F. H. Westheimer and J. E. Meyer, *J. Chem. Phys.* **14**, 733 (1946).
10. F. H. Westheimer, in M. S. Newman, eds., *Steric Effects in Organic Chemistry*, John Wiley & Sons, Inc., New York, 1956, p. 523.
11. J. B. Hendrickson, *J. Am. Chem. Soc.* **83**, 5537 (1961).
12. N. L. Allinger, J. Allinger, and M. A. DaRooge, *J. Am. Chem. Soc.* **86**, 4061 (1964) and references therein.
13. G. Herzberg, *Molecular Spectra and Molecular Structure. II. Infrared and Raman Spectra of Polyatomic Molecules*, Van Nostrand Reinhold, New York, 1945; E. B. Wilson, Jr., J. C. Decius, and P. C. Cross, *Molecular Vibrations*, McGraw-Hill Book Co., Inc., New York, 1955.
14. T. L. Hill, *J. Chem. Phys.* **14**, 465 (1946).
15. C. A. Coulson and G. S. Rushbrooke, *Proc. Cambridge Phil. Soc. Math. Phys. Sci.* **36**, 193 (1940).
16. M. Goepfert-Mayer and A. I. Sklar, *J. Chem. Phys.* **6**, 645 (1938).
17. M. J. S. Dewar, *J. Chem. Soc.*, 2329 (1950); M. J. S. Dewar, *J. Chem. Soc.*, 3532 (1952).
18. J. A. Pople, *Trans. Faraday Soc.* **49**, 1375 (1953); J. A. Pople, *J. Phys. Chem.* **61**, 6 (1957).
19. R. Pariser and R. Parr, *J. Chem. Phys.* **21**, 466, 767 (1953).
20. A. Streitwieser, Jr., *Molecular Orbital Theory for Organic Chemists*, John Wiley & Sons, Inc. New York, 1961.
21. J. A. Pople and D. L. Beveridge, *Approximate Molecular Orbital Theory*, McGraw-Hill Book Co., Inc., New York, 1970.
22. J. N. Murrell and A. J. Harget, *Semi-empirical Self-consistent-Field Molecular Orbital Theory of Molecules*, Wiley-Interscience, London, England, 1972.
23. R. Hoffman and R. B. Woodward, *Acc. Chem. Res.* **1**, 17 (1968); R. B. Woodward and R. Hoffman, *Conservation of Orbital Symmetry*, Verlag Chemie, Weinheim, Germany, 1970.
24. E. Wimmer, *Chem. Design Auto. News* **8**(4), 1 (1993); R. C. Merkle, *Chem. Design Auto. News* **8**(9–10), 1 (1993); see also *J. Comput.-Aided Mater. Design*.
25. A. L. Horvath, *Molecular Design*, Vol. 75, Elsevier, Amsterdam New York, 1992.
26. C. Levinthal, *Sci. Am.* **214**, 42 (1966).
27. R. Langridge and T. Klein, in C. A. Ramsden, vol. ed., *Comprehensive Medicinal Chemistry*, Vol. 4, *Quantitative Drug Design*, Pergamon Press, New York, 1990, pp. 413–430.
28. R. E. Hubbard, in N. C. Cohen, ed., *Guidebook on Molecular Modeling in Drug Design*, Academic Press, San Diego, Calif., 1996, pp. 19–54.
29. T. A. Jones, *J. Appl. Crystallogr.* **11**, 268 (1978).
30. D. Sayer, ed., *Computational Crystallography*, Oxford University Press, New York, 1982.
31. G. N. Phillips, Jr., *Biophys. J.* **69**, 1281 (1995).
32. R. Langridge, T. E. Ferrin, I. D. Kuntz, and M. L. Connolly, *Science* **211**, 661 (1981).
33. M. L. Connolly, *Science* **221**, 709 (1983); M. L. Connolly, *J. Appl. Crystallogr.* **16**, 548 (1983).
34. F. M. Richards, *Annu. Rev. Biophys. Bioeng.* **6**, 151–176 (1977).
35. K. Flurchik and L. Bartolotti, *J. Mol. Graphics* **13**, 10 (1995).
36. D. C. Richardson and J. S. Richardson, *Trends in Biochem. Sci.* **19**(3), 135–138 (1994).
37. N. Bohr, *Phil. Mag.* **26**, 476 (1913).
38. L. Pauling and E. B. Wilson, Jr., *Introduction to Quantum Mechanics*, McGraw-Hill Book Co., Inc., New York, 1935.
39. L. deBroglie, *The Revolution in Physics*, Noonday Press, Inc., New York, 1953; H. Eyring, J. Walter, and G. E. Kimball, *Quantum Chemistry*, John Wiley & Sons, Inc., New York, 1944; W. Kauzmann, *Quantum Chemistry*, Academic Press, New York, 1957.
40. P. A. M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, New York, 1958.
41. P. A. M. Dirac, *Proc. Roy. Soc. (London)* **A123**, 714 (1929).
42. A. Streitwieser, Jr., *Molecular Orbital Theory for Organic Chemists*, John Wiley & Sons, Inc., New York, 1961, pp. 6–7.
43. D. R. Hartree, *Proc. Cambridge Phil. Soc.* **24**, 89, 111, 426 (1927).
44. V. Fock, *Z. Phys.* **62**, 795 (1930).
45. A. Warshel and M. Karplus, *J. Am. Chem. Soc.* **94**, 5612 (1972).
46. N. L. Allinger and J. T. Sprague, *J. Am. Chem. Soc.* **95**, 3893 (1973).
47. J. A. Pople, D. P. Santry, and G. A. Segal, *J. Chem. Phys.* **43**, S129 (1965); J. A. Pople and G. A. Segal, *J. Chem. Phys.* **43**, S136 (1965).
48. M. Zerner, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. 2, VCH Publishers, New York, 1991, pp. 313–356.
49. R. C. Bingham, M. J. S. Dewar, and D. H. Lo, *J. Am. Chem. Soc.* **97**, 1285 (1975), and subsequent papers by Dewar and collaborators.
50. M. J. S. Dewar, E. F. Healy, J. J. P. Stewart, and W. Thiel, in Ref. 48.
51. J. J. P. Stewart, *J. Comput.-Aided Mol. Design* **4**, 1 (1990).
52. J. J. P. Stewart, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. 1, VCH Publishers, New York, 1990, pp. 45–82.
53. T. Clark, *A Handbook of Computational Chemistry*, John Wiley & Sons, Inc., New York, 1985, Chaps. 3 and 4.
54. D. Feller and E. R. Davidson, in Ref. 52, pp. 1–44.
55. W. J. Hehre, L. Radom, P. v.-R. Schleyer, and J. A. Pople, *Ab Initio Molecular Orbital Theory*, Wiley-Interscience, New York, 1986, and references therein.
56. C. E. Dykstra, J. D. Augspurger, B. Kirtman, and D. J. Malik, in Ref. 52, pp. 83–118.
57. D. B. Boyd, in Ref. 52, pp. 321–354.
58. M. J. S. Dewar, *J. Am. Chem. Soc.* **106**, 669 (1984).
59. L. S. Bartell, *J. Am. Chem. Soc.* **99**, 1977, 3279 (1977), and references therein.
60. U. Burkert and N. L. Allinger, *Molecular Mechanics*, ACS Monograph Series, Vol. 177, American Chemical Society, Washington, D.C., 1982, pp. 38, 157.
61. N. L. Allinger, *Adv. Phys. Org. Chem.* **13**, 1 (1976).
62. O. Ermer and S. Lifson, *J. Am. Chem. Soc.* **95**, 4121 (1973); O. Ermer, *Struct. Bonding* **27**, 161 (1976).
63. S. Lifson and A. Warshel, *J. Chem. Phys.* **49**, 5116 (1968); A. Warshel and S. Lifson, *Chem. Phys. Lett.* **4**, 255 (1969); A. Warshel and S. Lifson, *J. Chem. Phys.* **53**, 582 (1970); A. T. Hagler and S. Lifson, *Acta Crystallogr. Sect. B* **30**, 619 (1974); A. T. Hagler, E. Huler, and S. Lifson, *J. Am. Chem. Soc.* **96**, 5319 (1974); A. Warshel and M. Levitt, *Nature* **253**, 694 (1975). M. Levitt and A. Warshel, *J. Mol. Biol.* **106**, 421 (1976); M. Levitt and A. Warshel, *J. Am. Chem. Soc.* **100**, 2607 (1978).
64. D. H. Andrews, *Phys. Rev.* **36**, 544 (1930).
65. A. T. Hagler, L. Leiserowitz, and M. Tuval, *J. Am. Chem. Soc.* **98**, 4600 (1976); S. Lifson, A. T. Hagler, and P. Dauber, *J. Am. Chem. Soc.* **101**, 5111 (1979).
66. J. E. Williams, P. J. Stang, and P. v. R. Schleyer, *Annu. Rev. Phys. Chem.* **19**, 531 (1968).

67. K. B. Wiberg and R. H. Boyd, *J. Am. Chem. Soc.* **94**, 8426 (1972); K. B. Wiberg, *Comput. Chem.* **1**, 221 (1977).
68. S. Profeta, Jr., *Conformational Analysis of Aliphatic Amines by Quantum and Molecular Mechanical Methods*, Ph.D. Dissertation, University of Georgia, 1978, available from University Microfilms International, Ann Arbor, Mich.
69. Ref. 60, pp. 17–58.
70. Ref. 60, pp. 253–283.
71. S. Profeta, Jr. and N. L. Allinger, *J. Am. Chem. Soc.* **107**, 1907 (1985).
72. D. B. Boyd, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. 6, VCH Publishers, New York, 1995, p. 317–354.
73. U. Dinur and A. T. Hagler, in Ref. 48, pp. 99–164; J. R. Maple, U. Dinur, and A. T. Hagler, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 5350 (1988).
74. T. Halgren, *J. Comp. Chem.* **17**, 490, 520, 553, 587 (1996).
75. N. L. Allinger and co-workers, *J. Comp. Chem.* **17**, 642, 669, 695, 730, 747 (1996).
76. S. J. Weiner and co-workers, *J. Am. Chem. Soc.* **106**, 765 (1984); S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, *J. Comp. Chem.* **10**, 982 (1989).
77. J. A. McCammon and M. Karplus, *Annu. Rev. Phys. Chem.* **311**, 29 (1980); M. Karplus and J. A. McCammon, *Annu. Rev. Biochem.* **53**, 263 (1983); W. F. van Gunsteren and H. J. C. Berendsen, *Angew. Chem. Int. Ed. Engl.* **29**, 992 (1990).
78. C. R. Landis, D. M. Root, and T. Cleveland, in Ref. 72, pp. 73–148.
79. F. M. Menger and co-workers, *J. Am. Chem. Soc.* **100**, 4676 (1978).
80. E. K. Davies and N. W. Murrell, *Comput. Chem.* **13**, 149 (1989).
81. S. Profeta, Jr., R. J. Unwalla, B. T. Nguyen, and F. K. Cartledge, *J. Comp. Chem.* **7**, 528 (1986); S. Profeta, Jr., R. J. Unwalla, and F. K. Cartledge, *J. Org. Chem.* **51**, 1884 (1986); R. J. Unwalla, S. Profeta, Jr., and F. K. Cartledge, *J. Org. Chem.* **53**, 5658 (1988); S. Profeta, Jr., R. J. Unwalla, and F. K. Cartledge, *J. Comp. Chem.* **10**, 99 (1989).
82. M. R. Frierson, M. R. Imam, V. B. Zalkow, and N. L. Allinger, *J. Org. Chem.* **53**, 5248 (1988).
83. S. G. Cho, R. J. Unwalla, F. K. Cartledge, and S. Profeta, Jr., *J. Comp. Chem.* **10**, 832 (1989); S. G. Cho, F. K. Cartledge, R. J. Unwalla, and S. Profeta, Jr., *J. Mol. Struct.* **204**, 79 (1990); S. G. Cho, F. K. Cartledge, R. J. Unwalla, and S. Profeta, Jr., *Tetrahedron* **46**, 8005 (1990).
84. I. Pettersson and T. Liljefors, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. 9, VCH Publishers, New York, 1996, pp. 167–190.
85. D. B. Boyd, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vols. 1–7, VCH Publishers, New York, 1990–1995.
86. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, London, 1987.
87. J. M. Haile, *Molecular Dynamics Simulation: Elementary Methods*, Wiley-Interscience, New York, 1992.
88. C. L. Brooks III, M. Karplus, and B. M. Pettit, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*, John Wiley & Sons, New York, 1988.
89. J. A. McCammon and S. C. Harvey, *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, U.K., 1987.
90. W. van Gunsteren and P. K. Weiner, eds., *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, ESCOM, Leiden, The Netherlands, 1989.
91. I. Zuniga, I. Bahar, R. Dodge, and W. L. Mattice, *J. Chem. Phys.* **95**, 5348 (1991).
92. V. Galiatsatos, in Ref. 72, pp. 149–208.
93. W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.* **110**, 1657 (1988).
94. A. T. Hagler, in V. J. Hruby, ed., *The Peptides, Analysis, Synthesis and Biology: Conformation in Biology and Drug Design*, Vol. 7, Academic Press, Orlando, Fla., 1985, pp. 213–299.
95. W. L. Jorgensen and C. Ravimohan, *J. Chem. Phys.* **83**, 3050 (1985); W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
96. M. Saunders and co-workers, *J. Am. Chem. Soc.* **112**, (1990).
97. J. Skolnick and A. Kolinski, *Annu. Rev. Phys. Chem.* **40**, 207 (1989).
98. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
99. U. C. Singh and P. A. Kollman, Gaussian80 (UCSF), Quantum Chemistry Program Exchange, Prog. #446; *QCPE Bull.* **2**, 121 (1982); U. C. Singh and P. A. Kollman, *J. Comp. Chem.* **7**, 718 (1986).
100. P. A. Kollman and co-workers, AMBER: Assisted Model Building and Energy Refinement, University of California, San Francisco, 1980–present; S. J. Weiner and co-workers, *J. Am. Chem. Soc.* **106**, 765 (1984); S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, *J. Comp. Chem.* **10**, 982 (1989).
101. J. S. Binkley and co-workers, Gaussian 80 82, revision H, Carnegie-Mellon University, Pittsburgh, Pa., 1982.
102. J. Gao, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. 7, VCH Publishers, New York, 1996, pp. 119–186.
103. L. Bartolotti and K. Flurchick, in Ref. 97, pp. 187–216.
104. R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, New York, 1989.
105. SPARTAN, A Quantum Mechanical and Molecular Modeling Program Suite, available from Wavefunction, Inc., Irvine, Calif.
106. R. S. Pearlman, R. Balducci, A. Rusinko, J. M. Skell, and K. M. Smith, CONCORD: A Program for Generating Three-Dimensional Coordinates, available from Tripos Associates, St. Louis, Mo.
107. R. S. Pearlman, in H. Kubinyi, ed., *3-D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM Science Publishers, Leiden, the Netherlands, 1993, pp. 41–79.
108. A. Cayley, *Cambridge Math. J.* **II**, 267 (1841).
109. G. M. Crippen, *J. Comput. Phys.* **26**, 449 (1978).
110. G. M. Crippen and T. F. Havel, *Acta Crystallogr. Sect. A* **34**, 282 (1978); G. M. Crippen and T. F. Havel, *Distance Geometry and Molecular Conformation*, Chemometrics Research Studies Series 15, John Wiley & Sons, Inc., New York, 1988.
111. J. M. Blaney, G. M. Crippen, A. Dearing, and J. S. Dixon, DGEOM, Program #590 *QCPE Bull.* **10**, (1990), available from Quantum Chemistry Program Exchange, Indiana University, Bloomington, Ind.
112. T. F. Havel, I. D. Kuntz, and G. M. Crippen, *Bull. Math. Biol.* **45**, 665 (1983).
113. P. K. Weiner and co-workers, *Tetrahedron* **39**, 1113 (1983).
114. A. R. Leach, in Ref. 48, pp. 1–55.
115. J. M. Blaney and J. S. Dixon, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. 5, VCH Publishers, New York, 1994, pp. 299–335.
116. A. K. Ghose and G. M. Crippen, in Ref. 27, pp. 732–733.
117. QSPR: R. S. Pearlman, *Chem. Design Automat. News* **7**(3), 1 (1992); QSTR: V. K. Gombar, K. Enslein, and B. W. Blake, *Chemosphere* **31**, 2499 (1995); D. F. V. Lewis, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. 3, VCH Publishers, New York, 1992, pp. 173–222.
118. A. J. Leo and C. Hansch, *J. Org. Chem.* **36**, 1539 (1971); C. Hansch, in C. J. Cavallito, eds., *Structure–Activity Relationships*, Vol. 1, Pergamon Press, Oxford, U.K., 1973, p. 150; C. Hansch and A. J. Leo, *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington, D.C., 1995, Chap. 4 and 5.
119. J. M. McFadland, *J. Med. Chem.* **13**, 1192 (1970).
120. C. Hansch, A. R. Steward, S. M. Anderson, and D. Bentley, *J. Med. Chem.* **11**, 1 (1968); C. Hansch and J. M. Clayton, *J. Pharm. Sci.* **62**, 1 (1973).
121. Cerius², a Software Suite for Molecular Modeling, Version 3.5, Molecular Simulations, Inc., San Diego, Calif., 1997.

122. R. W. Taft, in M. S. Newman, ed., *Steric Effects in Organic Chemistry*, John Wiley & Sons, Inc., New York, 1956, pp. 556–675.
123. P. N. Craig, *J. Med. Chem.* **14**, 680 (1971).
124. C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.* **16**, 1212 (1973).
125. A. Leo, P. Jow, C. Silipi, and C. Hansch, *J. Med. Chem.* **18**, 865 (1975).
126. T. Fujita, in Ref. 27, pp. 497–560; C. Hansch and T. Fujita, eds., *Classical and Three-Dimensional QSAR in Agrochemistry*, ACS Symposium Series No. 606, American Chemical Society, Washington, D.C., 1995.
127. O. Kirino, H. Ohshita, T. Oishi, and T. Kato, *Agric. Biol. Chem.* **44**, 31 (1980).
128. C. D. Selassie, M. Chow, and C. Hansch, *Chem.-Biol. Interact.* **68**, 13 (1988).
129. Y. C. Martin, *Quantitative Drug Design*, Marcel Dekker, New York, 1978.
130. D. Cuthbert and F. S. Wood, *Fitting Equations to Data*, John Wiley & Sons, Inc., New York, 1971.
131. A. Cammarata, R. C. Allen, J. K. Seydel, and E. Wempe, *J. Pharm. Sci.* **59**, 1496 (1970).
132. J. G. Topliss and R. J. Costello, *J. Med. Chem.* **15**, 1066 (1972).
133. M. A. Pleiss and S. H. Unger, in Ref. 27, pp. 561–587.
134. H. Kubinyi, eds., *3-D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM Science Publishers, Leiden, The Netherlands, 1993.
135. A. Itai, Y. Mizutani, Y. Nishibata, and N. Tomioka, in N. C. Cohen, *Guidebook on Molecular Modeling in Drug Design*, Academic Press, San Diego, Calif., 1996, pp. 93–138.
136. P. S. Magee, D. R. Henry, and J. H. Block, *Probing Bioactive Mechanisms*, ACS Symposium Series No. 413, American Chemical Society, Washington, D.C., 1989.
137. C. H. Reynolds, M. K. Holloway and H. K. Cox, *Computer-Aided Molecular Design: Applications in Agrochemicals, Materials and Pharmaceuticals*, ACS Symposium Series No. 589, American Chemical Society, Washington, D.C., 1995.
138. R. D. Cramer, D. E. Patterson, and J. D. Bunce, *J. Am. Chem. Soc.* **110**, 5959 (1988).
139. S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, *SLAM J. Sci. Stat. Comput.* **5**, 735 (1984).
140. S. Wold, *Technometrics* **20**, 397 (1978).
141. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, and S. Clementi, *Quant. Struct.–Act. Relat.* **12**, 9 (1993); M. Baroni, D. Bonelli, S. Clementi, G. Cruciani, C. Ebert, and B. Skagerberg, *Quant. Struct.–Act. Relat.* **9**, 101 (1990).
142. K. G. Kowalski, *Chemometrics Intel. Lab Syst.* **9**, 177 (1990).
143. U. Thibaut, G. Folkers, G. Klebe, H. Kubinyi, A. Merz, and D. Rognan, in Ref. 134, pp. 711–728.
144. D. B. Boyd, in Ref. 102, pp. 303–380.

General References

Historical perspective: Ref. 3 is also a general reference.

R. M. Davis, *Science* **195**, 1099 (1977).

M. S. Tute, in C. A. Ramsden, vol. ed., *Comprehensive Medicinal Chemistry*, Vol. 4, *Quantitative Drug Design*, Pergamon Press, New York, 1990, pp. 1–32.

Definition of Molecular Modeling and Uses of CAMM:

L. M. Balbes, S. W. Maserella, and D. B. Boyd, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, VCH Publishers, New York, Vol. V, 1994, pp. 337–379.

Chemical Design Automation News, Elsevier Science, Inc., New York, 1985–1997.

N. C. Cohen, ed., *Guidebook on Molecular Modeling in Drug Design*, Academic Press, San Diego, Calif., 1996.

D. Liotta, ed., *Advances in Molecular Modeling*, Vols. 1–2, JAI Press, Inc., Greenwich, Conn., 1988–1990.

Computer Graphics in Molecular Modeling:

Refs. 27 and 28 are also general references.

G. R. Marshall and C. B. Naylor, in C. A. Ramsden, vol. ed., *Comprehensive Medicinal Chemistry*, Vol. 4, *Quantitative Drug Design*, Pergamon Press, New York, 1990, pp. 431–458.

J. M. Blaney and C. Hansch, in C. A. Ramsden, vol. ed., *Comprehensive Medicinal Chemistry*, Vol. 4, *Quantitative Drug Design*, Pergamon Press, New York, 1990, pp. 459–496.

C. Levinthal, *Sci. Amer.* **214**, 42 (1966).

Computational Methods for Molecular Modeling:

L. Pauling and E. B. Wilson, Jr., *Introduction to Quantum Mechanics*, McGraw-Hill Book Company, Inc., New York, 1935.

L. DeBroglie, *The Revolution in Physics*, Noonday Press, New York, 1953.

H. Goldstein, *Classical Mechanics*, Addison-Wesley Publishing Company, Reading, Mass., 1956.

A. Streitwieser, Jr., *Molecular Orbital Theory for Organic Chemists*, John Wiley & Sons, Inc., New York, 1961.

J. Hammer, *The Conceptual Development of Quantum Mechanics*, McGraw-Hill Book Co., Inc., New York, 1966.

J. P. Tollenaere, in N. C. Cohen, ed., *Guidebook on Molecular Modeling in Drug Design*, Academic Press, San Diego, Calif., 1996, pp. 337–356. Contains a superb glossary of terminology.

Molecular Mechanics and Molecular Dynamics:

U. Burkert and N. L. Allinger, *Molecular Mechanics*, ACS Monograph Series, Vol. 177, American Chemical Society, Washington, D.C., 1982.

D. B. Boyd and K. B. Lipkowitz, *J. Chem. Educ.* **59**, 269 (1982). The method and underlying philosophy of molecular mechanics.

K. Rasmussen, in G. Berthier and co-workers, eds., *Lecture Notes in Chemistry*, Vol. 27, *Potential Energy Functions in Conformational Analysis*, Springer-Verlag, Berlin, 1985.

M. Karplus and J. A. McCammon, *Sci. Amer.*, 42 (Apr. 1986).

P. Kollman, *Annu. Rev. Phys. Chem.* **38**, 303 (1987).

W. F. van Gunsteren and P. K. Weiner, eds., *Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications*, ESCOM, Leiden, the Netherlands, 1989.

T. P. Lybrand, in K. B. Lipkowitz and D. B. Boyd, eds., *Reviews in Computational Chemistry*, Vol. 1, VCH Publishers, New York, 1990, pp. 295–320.

Reviews in Computational Chemistry, K. B. Lipkowitz and D. B. Boyd, eds., Vol. 1–10, VCH Publishers, New York, 1990–1997.

Structure Generation, QSAR, and CoMFA Modeling Methods:

Reference 134 is also a general reference.

C. A. Ramsden, vol. ed., *Quantitative Drug Design*, Vol. 4, *Comprehensive Medicinal Chemistry*, Pergamon Press, New York, 1990.

P. Krosgaard-Larsen and H. Bongaard, eds., *A Textbook of Drug Design and Development*, Harwood Academic Publishers, Switzerland, 1991.

F. Sanz, J. Giraldo, and F. Manaut, *QSAR and Molecular Modeling: Concepts, Computational Methods and Biological Applications*, J. R. Prous Science Publishers,

Barcelona, Spain, 1995.
M. Kuchar, ed., *QSAR in the Design of Bioactive Compounds*, J. R. Prous Science Publishers, Barcelona, Spain, 1992.
J. G. Topliss, *Perspectives in Drug Discovery and Design* **1**, 253 (1993).

Salvatore Profeta, Jr.
Monsanto



MOLECULAR RECOGNITION

Receptor–Substrate-/Host–Guest-Chemistry

Molecular recognition is a central point. It may be said that without molecular recognition, there would be no life in this world. Vital biochemical processes such as enzyme action, molecular transport, genetic information, processing and protein assembly all involve molecular recognition as an essential action (1–3). Understanding of its principles is still a problem although first elucidation of the rules that govern molecular recognition dates back to the late nineteenth century (4). Strictly speaking, in 1894 Emil Fischer, a farsighted chemist, came up with his brilliant "lock-and-key" idea. In his famous paper (5) he proposed that enzyme and substrate can be compared to lock and key getting selectivity between molecules involved. Moreover, and this occurred even earlier, it was Paul Ehrlich who recognized that molecules do not act if they do not bind, thus, introducing the concept of receptor (6). Finally binding or fixation requires interaction, affinity between the partners that may be related to the idea of coordination introduced by Alfred Werner (7).

According to these basic concepts, molecular recognition implies complementary lock-and-key type fit between molecules. The lock is the molecular receptor and the key is the substrate that is recognized and selected to give a defined receptor–substrate complex, a coordination compound or a supermolecule. Hence molecular recognition is one of the three main pillars, fixation, coordination, and recognition, that lay foundation of what is now called supramolecular chemistry (8–11).

Behind this new direction of supramolecular chemistry, the chemistry beyond the molecule, is a highly interdisciplinary field of science covering the chemical, physical, and biological features of chemical species of greater complexity than molecules themselves that are held together and organized by means of intermolecular (nonbinding) interactions (12). The chemistry of molecular recognition is also the core of host–guest chemistry which is a sub-discipline or a particular aspect of supramolecular chemistry mostly involving inclusion and complex formation (13) (see also INCLUSION COMPOUNDS).

Principles of Receptor Design

Information Storage and Read Out. Picking up the thread of the introduction, molecular recognition is defined by the energy and the information involved in the binding and selection of substrates by a given receptor molecule that may also involve a specific function (14). Mere binding is not recognition, although it is often taken as such. Instead, one may say that recognition is binding with a purpose, like receptors are ligands with a purpose. It implies a pattern recognition process through a structurally well-defined set of intermolecular interactions. Molecular recognition, thus, deals with the molecular storage and supramolecular read out of molecular information (9).

Information may be stored in the architecture of the receptor, in its binding sites, and in the ligand layer surrounding the bound substrate such as specified in Table 1. It is read out at the rate of formation and dissociation of the receptor–substrate complex (14). The success of this approach to molecular recognition lies in establishing a precise complementarity between the associating partners, ie, optimal information content of a receptor with respect to a given substrate.

Table 1. Structural Parameters for Storage of Information in a Chemical Receptor

Receptor	Parameter
architecture	size
	shape
	connectivity
	cyclic order
	conformation
	chirality
	dynamics
binding sites	electronic properties (charge, polarity, polarisability, van der Waals attraction and repulsion)
	size
	shape
	number
	arrangement
	reactivity (protonizable, deprotonizable, reducible, oxidizable)
	thickness
surrounding ligand layer	overall polarity (lipophilic, hydrophilic)
	specific polarity (exo endo-lipo polarophilic)

Complementarity. To a first approximation, complementarity should take two forms (Fig. 1). Firstly, the shape and size of the receptor cavity must complement the form of the substrate. Secondly, there must be a chemical complementarity between the binding groups lining the interior of the cavity and the external chemical features of the substrate (15).

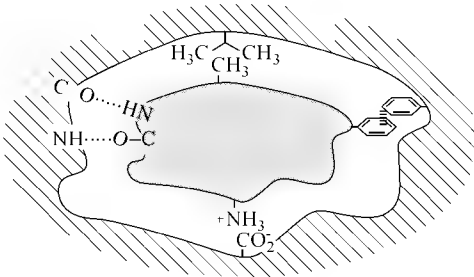


Fig. 1. Schematic representation of a receptor–substrate (host–guest) complex involving cavity inclusion of the substrate and the formation of different types of weak supramolecular interactions between receptor (hatched) and substrate (dotted).

The weak intermolecular forces that are principally involved in stabilizing receptor-substrate interactions and involved in molecular recognition processes (16) are summarized in Table 2. Examples are shown in Figure 1.

Table 2. Types of Interactions in Molecular Recognition

hydrogen bonding between basic and acidic centers
electrostatic attraction between anionic and cationic centers
metal-ligand interaction
dipole-dipole interaction
π -stacking and charge-transfer interaction between aromatic residues in the receptor and delocalized regions of the substrate
van der Waals attraction between hydrophobic regions on the two components
covalent bonds, that can be reversibly formed and broken (eg. disulfides, borate esters).

Most effective differentiation of the receptor between substrates will occur when multiple interactions are involved in the recognition process. The more binding regions (contact area) present, the stronger and more selective will be the recognition (17). This is the case for receptor molecules that contain intramolecular cavities, clefts or pockets into which the substrate may fit (Fig. 1).

Reorganization and Preorganization. On principle there are two different modes of receptor behavior illustrated in Figure 2. One of them, already evoked, is represented by the so-called lock-and-key image (Fig. 2a), involving complementary fit concept between rigid substrate and rigid receptor or rigid guest and rigid host relating to conformational flexibility of the molecular constituents forming the receptor-substrate (host-guest) complex (4). Receptors of this type are expected to present very efficient recognition between complementary partners, ie, both high stability and high selectivity of the receptor-substrate complex. The advantage of complementary receptor preorganization comes from minimizing the unfavorable entropy involved in substrate binding (see below).

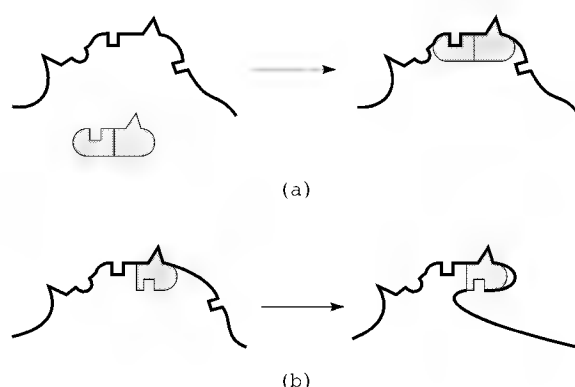


Fig. 2. Principle mechanisms of formation of a receptor-substrate complex: (a) Fischer's rigid "lock-and-key" model; (b) "induced fit" model showing conformational changes of the receptor (solid line) upon substrate binding.

However, in most biological system there is a degree of flexibility in the receptor (2). The approach of the substrate leads to conformational changes and an organization of the binding site around it. With this induced fit mechanism of binding (Fig. 2b), a higher entropy price is paid but there are several advantages (18). A flexible receptor will permit a more wrap around interception or even complete encapsulation with the substrate involving many more potential binding interactions. This may lead to high selectivity of binding involving the amplification of molecular recognition interactions illustrated in Figure 3 (19).

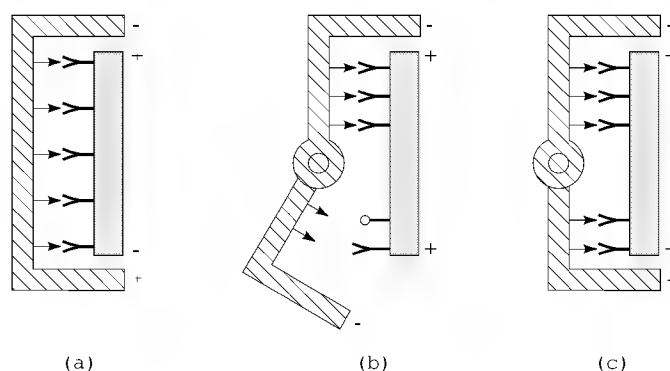


Fig. 3. Schematic approach illustrating amplification of molecular recognition effects of (a) matched rigid, (b) mismatched rigid, and (c) flexible type of receptor-substrate assemblies (19).

In case of the rigid lock-and-key type receptor forming five hydrogen bonds plus two extended electrostatic attractions (Fig. 3a), one mismatched hydrogen bond will result in only a small reduction in overall binding free energy $4.18\text{--}8.36\text{ kJ mol}^{-1}$ ($1\text{--}2\text{ kcal mol}^{-1}$) out of $\sim 41.8\text{ kJ mol}^{-1}$ ($\sim 10\text{ kcal mol}^{-1}$). The small difference in association constants (K) that would result ($\sim$ two orders of magnitude) is not sufficient, eg, for the differentiation of chemically similar substrates commonly encountered with biological systems (1). One solution would be using a flexible version of the receptor (Fig. 3b) that will profit from the amplification of molecular recognition interaction. Here a single mismatched hydrogen bond due to a repulsive interaction forces the receptor into a nonproductive binding conformation leaving two hydrogen bonding and an electrostatic site distant from the substrate. The difference in binding is, thus, amplified from a single hydrogen bond to almost half the binding interactions relative to the matched case in Figure 3c, which is profitable for selectivity.

However, reduced stability of the receptor-substrate complex as such involving a flexible receptor is the other side of the coin, since part of the binding energy is used up in the change of conformation of the receptor. This is the point amounting to the so-called principle of preorganization which can be expressed as follows: "the smaller the changes in organization of receptor and substrate or host and guest required for complex formation, the stronger is the binding" (20). An illustration of this approach is given in Figure 4 considering a systematic series of well-known synthetic receptors (crown

compounds and analogues, see also [INCLUSION COMPOUNDS](#)) that owe their binding properties to varying degrees of preorganization, ie, organization of their binding sites (donor atoms) prior to complexation (21).

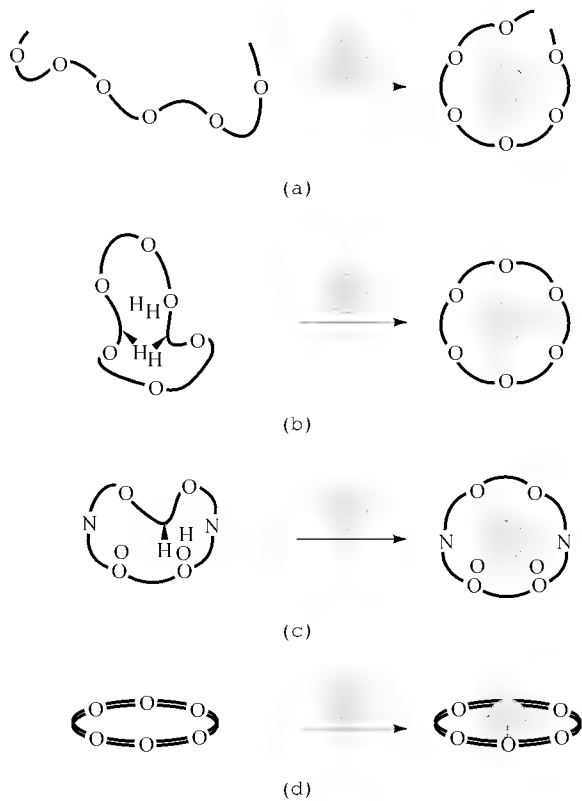


Fig. 4. Recognition and preorganization of hosts (receptors) on complex formation (21).

Podands (Fig. 4a) are acyclic collections of binding sites held together by appropriate spacer units (22,23). During the complexation act to form a podate (podaplex), many degrees of conformational freedom must be frozen out. Crowns and coronands (Fig. 4b) are cyclic collections of binding sites and are less flexible than podands (13,24). Nevertheless, they possess a variety of conformations, many of which fill their own potential cavities with their own spacer units. Cryptands (Fig. 5c), which are bridged crown analogues, naturally have a smaller number of nonbinding conformations, and therefore, require less expenditure of reorganization on inclusion formation than crowns (25,26). The so-called spherands (27,28) (Fig. 4d) are at the end of the unfolded progression of receptor structures with regard to the parameter of preorganization (20,21). Hence they are characterized as completely preorganized receptor systems possessing an enforced cavity with perfect octahedral arrangement of six donor oxygens and cavity diameter to be complementary to Li^+ and Na^+ in both an electronic and steric sense, giving rise to extremely high stability constants for complexes (spheraplexes) with these cations ($K_s\text{Li}^+ > 7 \cdot 10^{16}$, $K_s\text{Na}^+ = 1.2 \cdot 10^{14}$; in CDCl_3 sat. with D_2O at 25 °C) unlike K^+ and larger cations (cf Table 3) that are not complexed (20,28). From the kinetic point of view the facts are different and the order is reverse, ie, the rigid highly preorganized spherands are slow, as contrasted with the flexible barely preorganized podands that are fast both in formation and decomposition of the receptor–substrate (host–guest) complex (20,21).

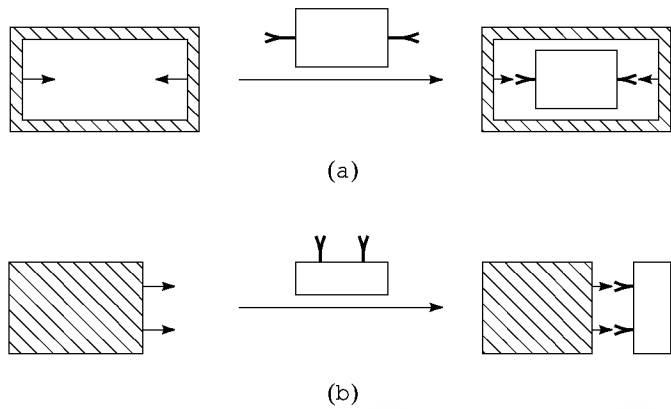


Fig. 5. Diagram of (a) endo- and (b) exo-receptor recognition of substrates (20).

Table 3. Comparison of Cation and Cavity Diameters

Cation	Cation diameter, ^a E	Crown ether	Cavity diameter, ^a E
Li^+	1.36	12-crown-4 (1a)	1.2–1.5
Mg^{2+}	1.56		
Na^+	1.90	15-crown-5 (1b)	1.7–2.2
Ca^{2+}	2.12		
Sr^{2+}	2.54	18-crown-6 (1c)	2.6–3.2
K^+	2.66		
Ba^{2+}	2.86		
Rb^+	2.98		
Cs^+	3.38	21-crown-7 (1d)	3.4–4.3

^a — 0.1 nm.

Hence, the balance between rigidity and flexibility is of particular importance for the binding and the dynamic properties of a receptor. It is, thus, a decisive structural design parameter of the receptor depending on the use. For instance, processes of exchange regulation, cooperativity and allostery connected with molecular recognition require a built-in flexibility so that the receptor may adapt and respond to changes unlike rigid receptors (14,29).

Topology. This parameter may have reference to either the receptor as an individual molecular structure or to the receptor–substrate complex on a higher level of organization that is directly related to the mode and efficiency of molecular recognition (14,30).

It has already been stressed that a concave receptor is a favorable case. Under these circumstances the receptor cavity is lined with binding sites directed toward the bound species (see Fig. 1). This corresponds to Cram's definition of a receptor (host) molecule providing binding sites that are convergent, as contrasted with the bound substrate (guest) featuring divergent complementary sites, ie, the substrate is more or less completely surrounded by the receptor forming an inclusion complex (20,31). This widely used principle of convergence defines a convergent or endo-supramolecular chemistry (host–guest chemistry) with endo-receptors (endo-hosts) effecting endo-recognition (Fig. 5a) (9).

The opposite procedure consists in making use of an external receptor surface rather than an internal cavity as substrate receiving site. This amounts to the passage from a convergent endo-supramolecular chemistry to a divergent or exo-supramolecular chemistry, and from endo- to exo-receptors (Fig. 5b) (9). Here receptor–substrate binding occurs by surface-to-surface interaction which may be termed affixation as contrasted with inclusion. Exo-recognition with strong and selective binding, in particular, requires a large enough contact area and a sufficient number of complementary interactions along the interface. Such a mode of molecular recognition also finds biological analogies, for instance at the antibody–antigen interface of immunological importance (32). Metallo-exoreceptor aggregation, molecular recognition at organic and inorganic monolayers, films and solid surfaces bearing recognition groups, as well as the design of supramolecular solid architectures and materials, are other important instances of the exo-recognition principle discussed in more detail below.

Apart from the basic classification in convergent endo- and divergent exo-receptors, molecular receptors (hosts) of extremely varied structural types have been developed (30) including the acyclic podands, the macrocyclic crown ethers, coronands or torands, the macropolycyclic cryptands and speleands, the spherands, cavitands and carcerands, the calixarenes and cryptophanes, the clathrands, the helicands and other organization framework species (12). Each of these trivial names refers to a particular aspect of the structure involving the overall receptor topology, that is connectivity, dimensionality and cyclic order. Examples of compounds are illustrated elsewhere in this article (see also INCLUSION COMPOUNDS) while a selection of possible topologies are shown in Figure 6, from a linear receptor (a) to spherical (i) or cylindrical (j) tricyclic structures following the classification of graphs (14). Moreover, all these types of receptors may possess a single receptor unit to recognize and bind a single substrate, eg, (d) or contain more than one discrete binding subunit (h) being characteristic of monotopic and oligo- polytopic receptors, respectively. The size and the shape of the binding cavity that they define and their rigidity or flexibility are determined by the nature of the structural subunits making up the branches of the graphical representations in Figure 6. In this respect they serve the same purpose for conveying properties of molecular recognition, mostly based on geometrical discrimination.

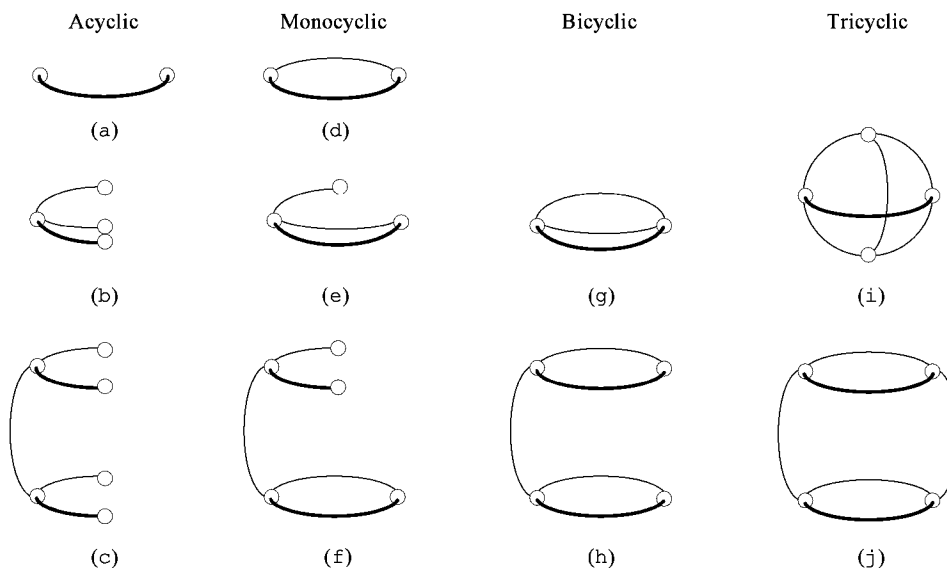


Fig. 6. Topological structure types of receptors (hosts) (14).

Simple Modes of Molecular Recognition

Substrates involved in molecular recognition may feature a particular shape, size, state of charge, chemical affinity or optical specification (19,30,33–36). In general most of these parameters share. Nevertheless there may be dominating features of a certain substrate molecule to be used by a complementary receptor in the recognition process (9).

Size and Shape Dominated Substrate Recognition. Perhaps the simplest recognition process is that of a spherical substrate, in its most elementary form a ball-shaped metal ion of defined diameter. Supramolecular chemistry in itself started with this problem, in particular having the effort for recognition and binding of alkali and alkaline-earth cations (37). Numerous studies have been performed that are reported in many papers and summarized in reviews and books (12) showing that three main classes of receptors providing spherical recognition property may be distinguished (38). They are (1) macrocyclic polyethers, the well-known crown ethers and their derivatives (24); (2) the macropolycyclic cryptands (25,26); and (3) the acyclic analogues of crown compounds and cryptands usually designated as podands (22,23). Prototypical compounds for each substance class are given by compounds (1)–(3) (Fig. 7). They all possess a spherical or quasispherical negatively polarized cavity prepared for the accommodation of alkali- and alkaline-earth metal ions that have complementary size, giving rise to a feature termed as spherical recognition (9). Coronates, cryptates or podates are the names of the respective inclusion complexes (39).

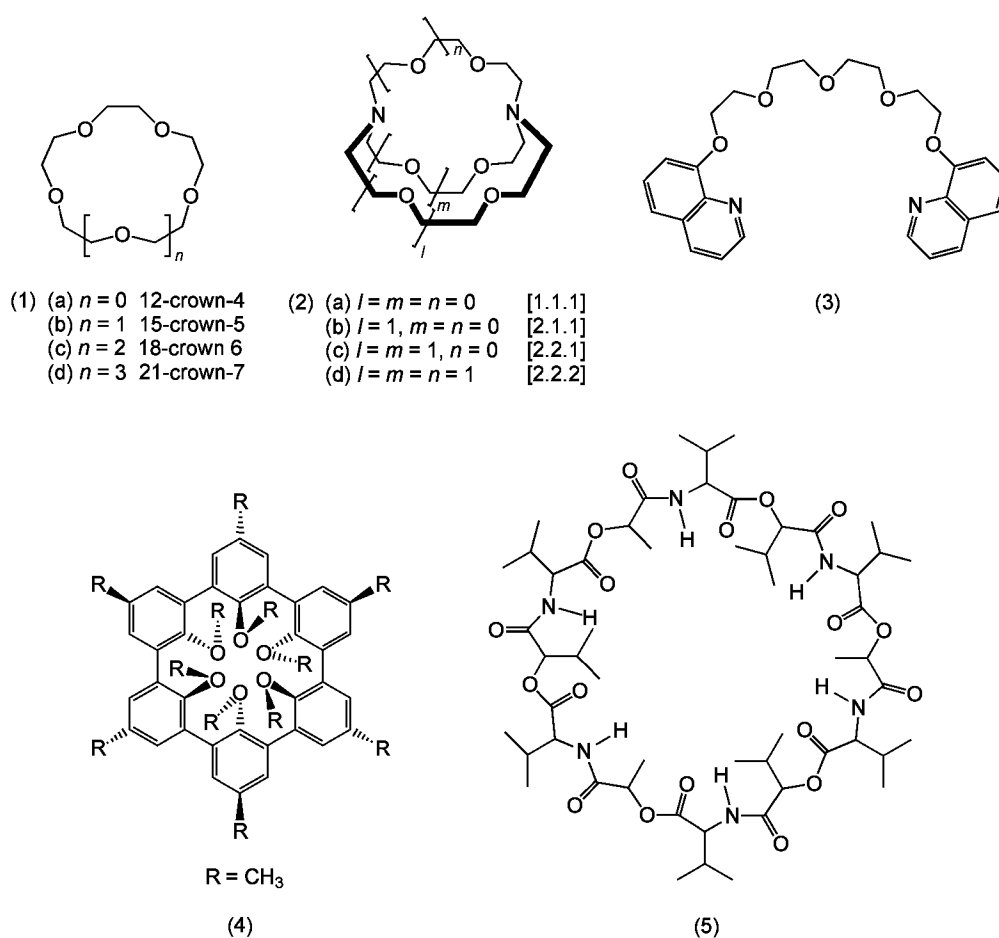


Fig. 7. Crown type and analogous receptor molecules of different varieties; (1) crown ethers; (2) cryptands; (3) a podand; (4) a spherand; and (5) the natural depsipeptide valinomycin.

The match between crown cavity diameter and cation diameter is obvious from Table 3 showing that, eg, Li^+ and 12-crown-4 (**1a**) or K^+ , respectively Ba^{2+} and 18-crown-6 (**1c**) correspond. Similar are the cryptands of gradually increasing cavity size [2.1.1], [2.2.1] and [2.2.2] for Li^+ , Na^+ and Sr^{2+} or K^+ and Ba^{2+} , while the small cavity of [1.1.1] fits H^+ . The example of such a matched cryptate where K^+ is accommodated into the cavity of [2.2.2] is illustrated in Figure 8a (25). Although acyclic podands do not provide a permanent cavity, they may create one by encircling a spherical cation with the length of the receptor molecular thread being the controlling parameter (22,40). Nevertheless, from what has already been said, low preorganization and topology of the podands handicap the substrate recognition which is increasingly higher in the circular crown and spheroidal cryptand case, but is most pronounced for the spherand type of receptor (eg, **4**) also mentioned before.

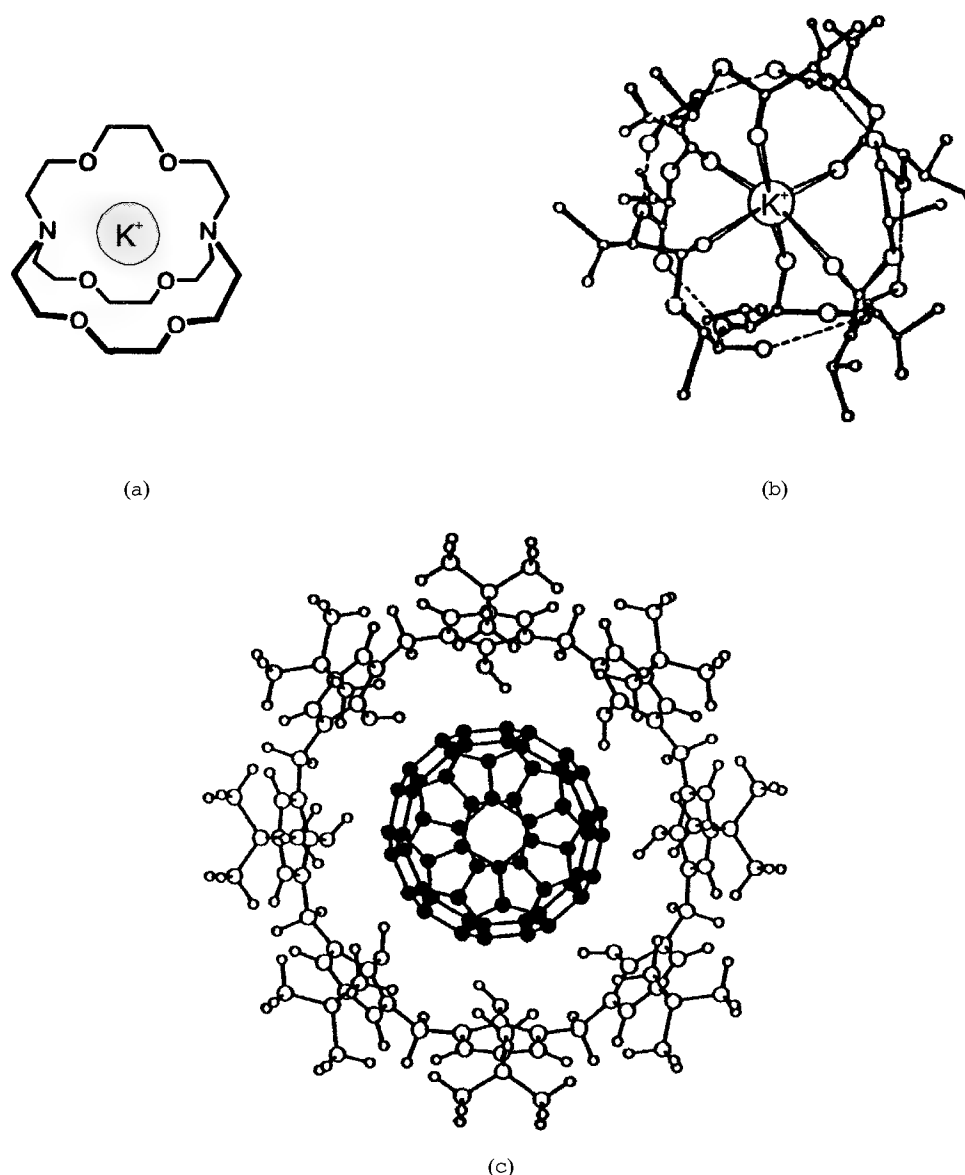


Fig. 8. Spherical recognition: (a) K^+ cryptate of [2.2.2] (**2d**); (b) K^+ complex of valinomycin (**5**); and (c) inclusion compound of C₆₀ into *tert*-butylcalix[8]arene.

Alkali and alkaline-earth cation recognition is also affected rather efficiently by other macrocyclic receptors that have been synthesized (21) such as the cryptospherands (an amalgamation of cryptand and spherand topology) (20,31) and lariat ethers (41), the latter is characterized by a crown ether ring that has attached extra podand arms for feeling and lock in of the substrate. Moreover, natural macrocycles displaying antibiotic properties are also very efficient in the recognition of alkali metal ions (42). For instance, valinomycin (**5**) gives a strong and selective complex in which a K^+ ion is included in the macrocyclic cavity in octahedral environment of six carbonyl oxygens (Fig. 8b) (43).

In a word, all these receptors are more or less able to discriminate against cations that are either smaller or larger than their cavity (44). However, in a strict sense, discrimination of metal-ion spheres does not concern with molecular recognition but selection of the carbon ball C₆₀ certainly does. In fact, the fullerene C₆₀ has been included into the cavity of octa-*tert*-butylcalix[8]arene (Fig. 8c) shutting out C₇₀ and making a very convenient and efficient C₆₀ purification possible without any expensive apparatus (45).

Recognition of a tetrahedral substrate geometry requires the construction of a receptor molecule with a tetrahedral recognition site (9). This may be realized by positioning four suitable binding sites at the corners of a tetrahedron and incorporating them into a bridged molecular framework such as shown with compound (**6**) (Fig. 9a) (25,38,46). In fact, the tetrahedral NH_4^+ cation is very firmly held inside the cavity of the tricyclic cryptand (**6**), forming the ammonium cryptate (Fig. 9b) (47). This cryptate presents a high degree of receptor–substrate complementarity in that the ammonium ion fits into the cavity of (**6**) and is held by a tetrahedral array of hydrogen bonds including also electrostatic interactions with the oxygen atoms (48). Unsymmetrical derivatives of (**6**) display notably perturbed NH_4^+ binding, with a marked loss in recognition behavior (49).

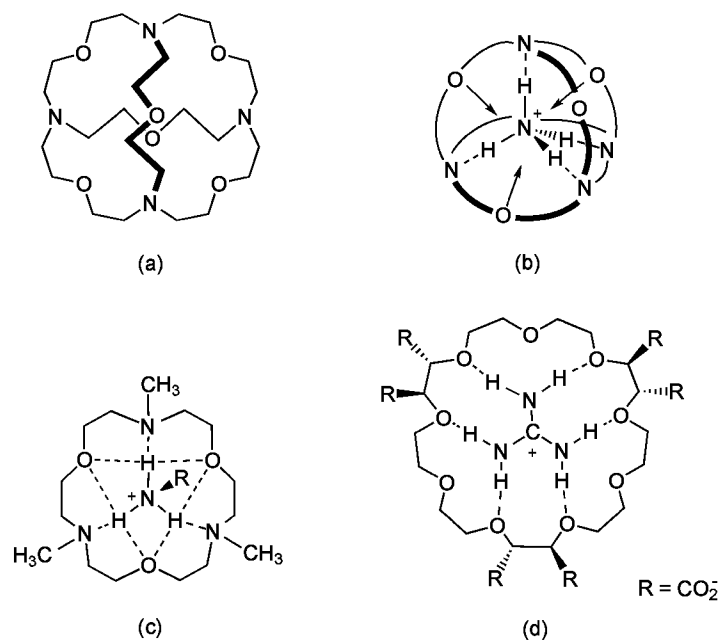


Fig. 9. Representative examples of 6 (a, b) tetrahedral; (c) trigonal; and (d) circular recognition.

Recognition of a primary ammonium ion, by analogy, is achieved by making use of a symmetrical triazacoronand enabling a trigonal recognition process in its inclusion complex (Fig. 9c) (50). A recognition site for secondary ammonium groups is provided by the diaza analogue of 12-crown-4 [cf (1a)] (51). It binds via two hydrogen bonds (52). On the other hand, the derivative of a 27-membered crown compound yielded a particular stable inclusion complex with the guanidinium cation that is bound through an array of six hydrogen bonds suggesting an almost circular recognition mode (Fig. 9d) (53,54).

Linear recognition is mainly a subject of molecular length recognition (9). For that reason preferential substrates bear two recognizable functional groups at a distance corresponding to a ditopic receptor molecule (Fig. 10a) (55). An example is given by the cylindrical macrotricyclic cryptand that yields cryptates with terminal diammonium cations $H_3N^+-(CH_2)_n-N^+H_3$ of matching molecular length (Fig. 10c) (56). A complementary substrate, eg, relating to the receptor with R = naphthalene-2,6-diyl in Figure 10c, would be the 1,5-pentane diammonium cation $[A = (CH_2)_5]$. In the resulting complex, the substrate is located in the central cavity and anchored by its two NH_3^+ groups in the macrocyclic binding sites (57).

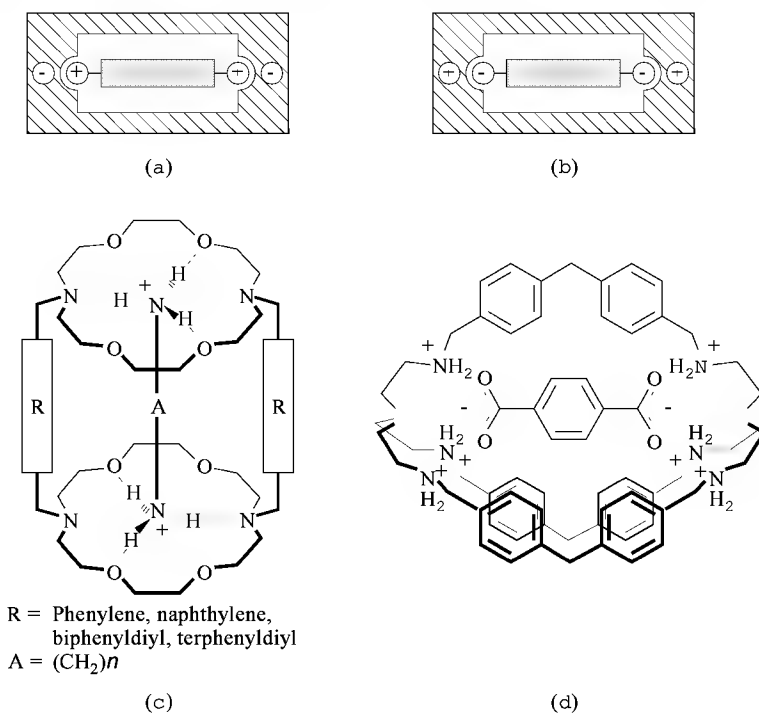


Fig. 10. Linear recognition: diagrammatic representation of the recognition of linear dicationic (a) and dianionic (b) substrates; (c, d) typical examples of receptor-substrate complexes.

Charge Attraction Dominated Recognition. Thus far, for recognition sizes and shapes of the substrate have been the focus. Nevertheless, charge attraction between the substrate and the receptor has also played a part since cations such as metal ions or ammonium ions were complexed by negatively polarized cavities. But metal ions involved hard alkali and alkaline-earth cations rather than weaker transition metal ions. Replacing the oxygen sites of crown compounds and cryptands with nitrogen (58) and sulfur atoms (59,60) yields receptors [eg, (7), Fig. 11] that show marked preference for transition-metal ions (61) and may allow high selective recognition of toxic heavy-metal ions such as cadmium, lead or mercury according to the hard and soft acid and base (HSAB) principle (62). Others containing internally directed functionalized units such as in the triscatechol derivative (8, Fig. 11) form very strong and selective complexes with Fe³⁺ or actinide and lanthanide ions (63,64) while a similar receptor with hard endocarboxylic acid groups is efficient for hard Ca²⁺ and Mg²⁺ ions showing again responsibility of a charge density effect in the receptor-substrate recognition (65). Thus, recognition of hard alkali and alkaline-earth metal ions is determined by coulombic attraction whereas the weak transition-metal ions are mainly controlled by geometrical parameters of orbital overlap.

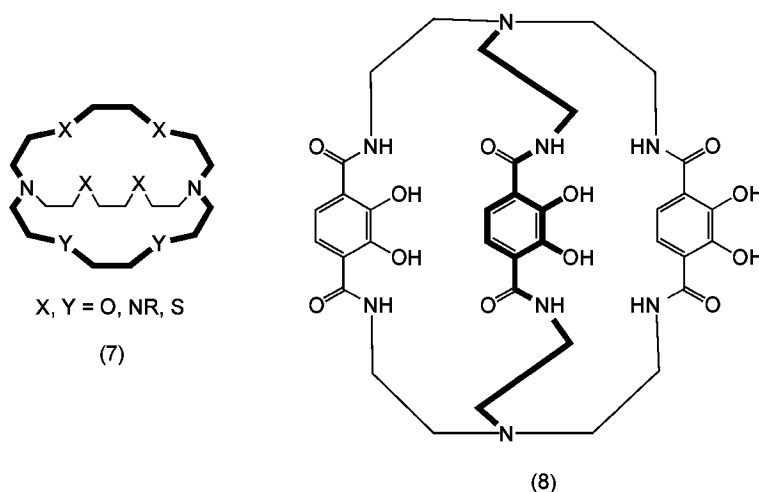


Fig. 11. Receptor molecules (cryptands) having hetero (nonoxygen) donor atoms (7) or endo-functional acidic sites (8) in the framework.

By analogy, recognition of an anionic substrate requires an electron-deficient receptor cavity including positively charged or neutral electron-deficient groups that may serve as interaction sites for anion binding (9,30). Ammonium and guanidinium units, which form $N^+ - H \cdots X^-$ bonds, have mainly been used (66), but electron-deficient boron, tin, mercury or metal ion centers in complexes also interact with anions (67). Apart from this demand, anionic substrates have another specific feature in that they are large compared with cations. Moreover they possess a range of geometries: spherical (eg, halides), linear (eg, N_3^- , CN^- , OCN^-), planar (eg, NO_3^- , $RCOO^-$) or tetrahedral (eg, SO_4^{2-} , ClO_4^- , phosphates). Polyammonium macropolycycles have been studied most extensively as receptor molecules for anion recognition (68). When (6) (Fig. 9a) is tetraprotonated it binds Cl^- very strongly and very selectively compared with Br^- and other types of anions giving the size complementary chloride cryptate, in which the included anion is bound by four $N^+ - H \cdots Cl^-$ hydrogen bonds (Fig. 12a) making spherical recognition of the chloride ion possible (69). Quaternary ammonium derivatives of oxygen-free macrotricycles of type (6) bind spherical anions as well (70), and more recently acyclic quaternary polypyridinium and polybipyridinium receptors for chloride and bromide recognition have also been reported (71). Linear recognition is displayed by the hexaprotonated form of ellipsoidal bis-tren type cryptands which bind with length discrimination various polyatomic anions and extend the recognition of anionic substrates beyond the spherical halides such as triatomic anion N_3^- (72) or dicarboxylate ions (73) via a pattern schematically shown in Figure 10b or illustrated for a case of dicarboxylate recognition in Figure 10d. Anion recognition has also made significant progress by means of a receptor type that features a guanidinium group (66). This particular group is of special interest for the recognition and binding of carboxylate and phosphate functions and related species since it may form two chelating H-bonds with the anionic units. An example is given in Figure 12b (74).

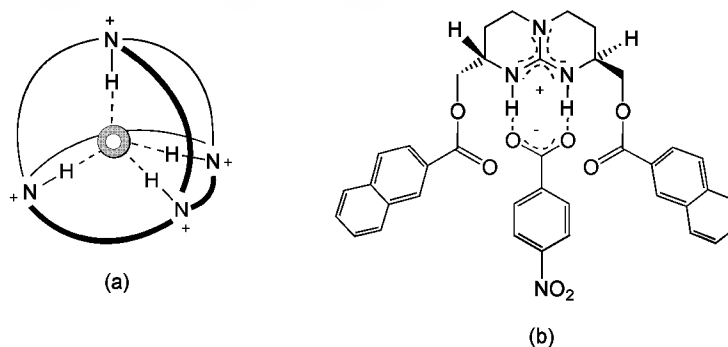


Fig. 12. (a) Spherical anion (Cl^-); and (b) carboxylate anion recognition.

Hydrogen Bond Dominated Recognition. Recognition of bioactive compounds is largely determined by the use of hydrogen bonding between polar sites (75). Here substrate recognition results from the formation of specific hydrogen bonding pattern between complementary subunits, in a way reminiscent of base pairing in nucleic acids (76). Hence hydrogen bonding has also been determined an important parameter for the design of artificial receptors (19,30,35,77,78). In fact, a great many of the above receptors and substrates where ammonium and anionic groups are involved base on this interaction mode. A more simplified version of hydrogen bond dominated receptors is seen with 2-aminopyridine derivatives, mostly of amide type (19). They can form specific hydrogen bonds to carboxylic acids such as illustrated in Figure 13a showing a respective receptor substrate complex between a picoline diamide receptor and complementary glutaric acid substrate that binds selectively via a length discrimination against dicarboxylic acids of larger and smaller chain length (79). Vice versa, dicarboxylic acid-type receptors are also efficient in the recognition of 2-aminopyridines, respectively of 2-aminopyrimidine (Fig. 13b) (80).

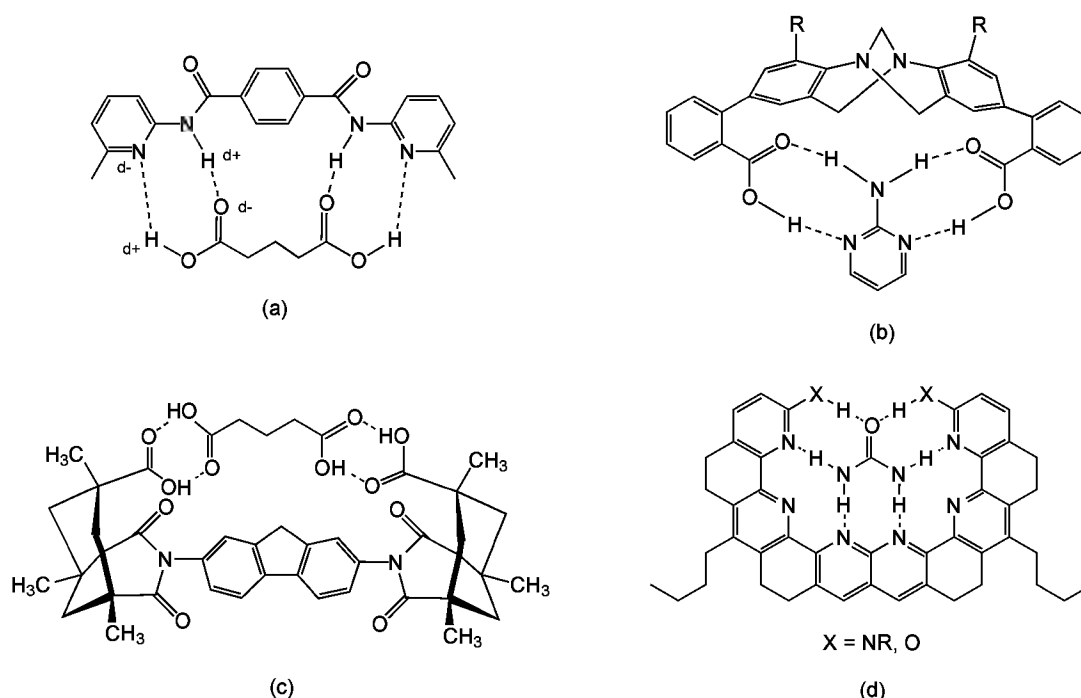


Fig. 13. Hydrogen bond dominated substrate recognition of: (a), (c) dicarboxylic acids; (b) 2-aminopyrimidine; and (d) urea.

An alternative approach to dicarboxylic acid recognition has been developed by using a receptor cleft based on a rigid spacer and two Kemp's triacid binding sites specified in Figure 13c (81–83). In this case, the principal binding force is a double carboxylic acid dimer interaction leading to both strong and selective binding to carboxylic acids of complementary shape and size, ie, glutaric acid (84). Another obvious cleft-type receptor showing both a rigid molecular framework and complementary orientation of hydrogen bonding (85) to allow, eg, the efficient recognition and binding of urea is illustrated in Figure 13d (86). Although the receptor clefts that use carefully designed and rigid components to hold hydrogen bonding groups at a fixed distance apart are very common in this field, macropolycycles comprising similar functions offer an alternative albeit more synthetically challenging solution to the same problem, discussed more detailed below where the recognition of particular substrate molecules is dealt with.

π -Stacking and Charge-Transfer Dominated Substrate Recognition. Nature's strategy for the recognition of substrates featuring a flat aromatic framework (planar recognition) affords another recognition element, namely π – π stacking interactions between aromatic rings, ie, aromatic groups of receptor and substrate that meet a parallel face-to-face orientation, apart from hydrogen bonding being also typical of the nucleotide recognition (77). A very simple example showing the principles of π – π stacking supported substrate recognition is illustrated in Figure 14a (87). The flat heterocyclic substrate, uracil derivative, fits in face-to-face mode into the conformationally stepped receptor macrocoring containing a naphthalene π -stacking unit and being bound via a system of extra hydrogen bonds to a diamidopyridine unit. The association constant for the analogous receptor molecule without the π -stacking unit is more than four times lower. The role of the π -stacking subunit in the above receptor clearly indicates that doubling the π -stacking contribution should lead to substantial improvement of recognition behavior and increase of binding energy. This has led to a molecular tweezer design strategy to flank a hydrogen bonding carboxylic acid by two π -stacking anthracene units corresponding to the receptor–substrate relationship demonstrated in Figure 14b (88).

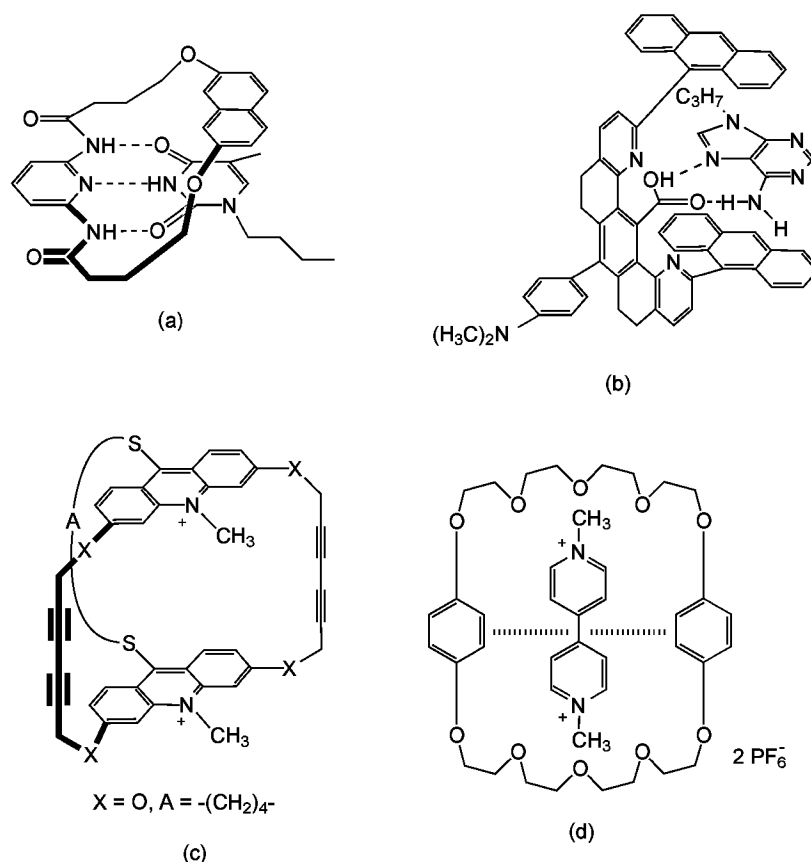


Fig. 14. π -Stacking and charge-transfer dominated recognition of flat aromatic–heteroaromatic substrates (formation of intercalates).

In a way, π - π stacking and charge transfer type of recognition have something in common. Examples making facts obvious are presented in Figure 14c and 14d. The macrobicyclic intercaland of Figure 14c and related receptors have been found to recognize flat shaped substrates through π - π stacking and bind them to form a molecular cryptate, in particular if electron donating substrate species are involved to allow charge-transfer interaction, such as planar molecular anions or nucleic acids (89). Similarly, large ring electron donor aromatic crown ethers were designed that yield charge-transfer type intercalation complexes with paraquats (Fig. 14d) (90). Vica versa, macrocyclic bipyridinoquates form charge-transfer supported intercalation complexes with flat aromatic substrates having electron-donating substituents (91).

Lipophilic Interaction Dominated Substrate Recognition. Making recognition through lipophilic interaction possible require receptors presenting large and more or less rigidly connected architectures of macrocyclic or cage-like nature (92). Here only some illustrative examples can be given (see also INCLUSION COMPOUNDS), referring the reader to specific reviews of this vast subject (8–12).

The naturally occurring cyclodextrins having endo-lipophilic cone-shape are perhaps the most important and also the first receptor molecules whose selective inclusion properties toward lipophilic organic molecules were recognized (93,94). They comprise a family of cyclic oligosaccharides, composed of 6, 7, and 8 glucose units in its most familiar representatives (α , β , and γ -cyclodextrin, respectively) providing endo-lipophilic and exo-hydrophilic cone-shaped molecular cylinders of increasing size (Fig. 15a). Cyclodextrins form size and shape selective inclusion compounds with a wide variety of substrates including benzene derivatives, paraffins, and noble gases (95).

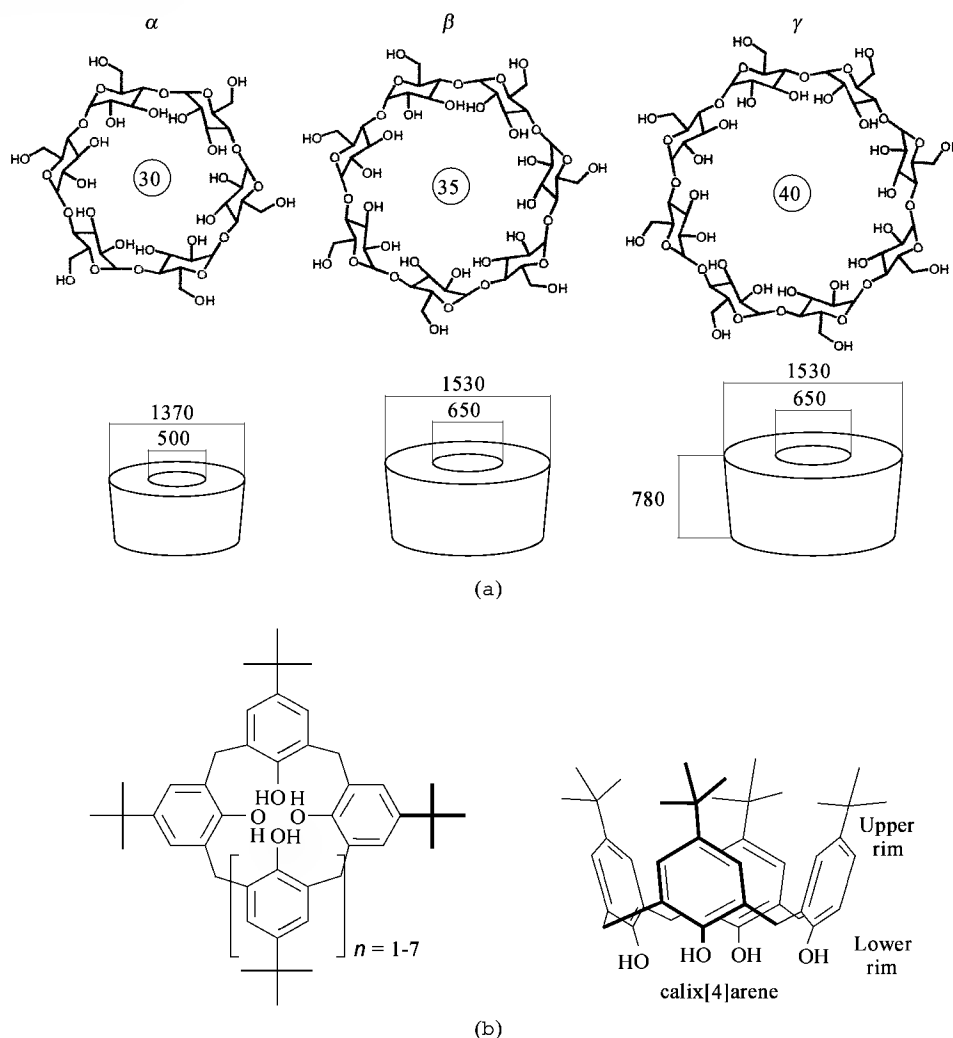


Fig. 15. Prototype examples of (a) cyclodextrins and (b) calixarenes, showing conformational structures and dimensions.

Calixarenes (from the Latin *calix*) may be understood as artificial receptor analogues of the natural cyclodextrins (96,97). In its prototypical form they feature a macrocyclic metacyclophane framework bearing protonizable hydroxy groups made from condensation of *p*-substituted phenols with formaldehyde (Fig. 15b). Dependent on the ring size, benzene derivatives are the substrates most commonly included into the calix cavity (98), but other interesting substrates such as C_60 have also been accommodated (Fig. 8c) (45).

As mentioned, calixarenes fall into the cyclophane-type of compounds that has emerged the central class of synthetic receptors in molecular recognition involving all kinds of ortho-, meta- and para-bridged aromatic macrocycles and oligomacrocycles, ie, pocket, open vessel, and macrocage receptors (99,100). A beautiful construction of this type are the cryptophanes, an example of which is shown in Fig. 16a (101,102). Cryptophanes are of much interest in particular for their ability to recognize and bind derivatives of methane that match the cavity. Although substrates are rather cut off outside here, the most extreme case of imprisonment of substrates is provided by the carcerplexes (103). They are the inclusion complexes of carcerands (104). A prototype example is illustrated in Figure 16b. These containers have a virtually closed molecular surface indicating that carcerplexes are formed during shell closure of two hemisphere components (cavitands) templated around the substrate (31,105,106). They also found that high structural recognition is involved in this capture, and the shell closure to give empty carcerands do not occur. Once substrates have been trapped in the cavity they can be released again only by destruction of the carcerand framework. Hemicarcerands are like carcerands except that by conformational modifications they can generate portals which join the inner phase of the receptor with the outer phase making an equilibrium of the substrate between in and out possible (103–106).

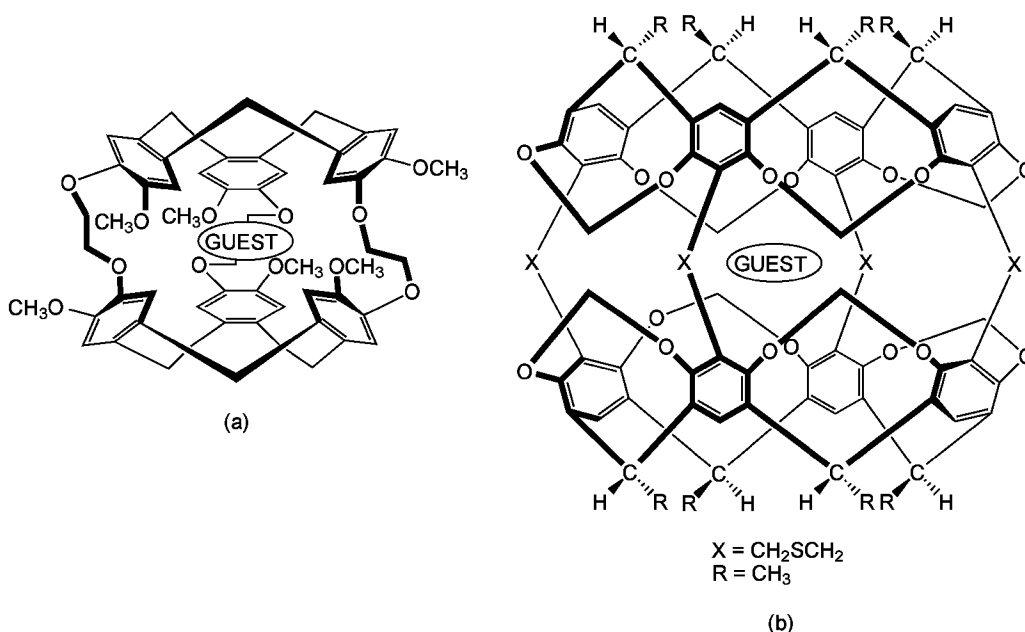


Fig. 16. (a) Inclusion complex of a cryptophane; and (b) a carceplex (carcerand inclusion complex).

While the previous receptors are typically used in organic solvents, except for the cyclodextrins, there are special cases of cyclophane receptors supplied with peripheral charges (ammonium units) (107–12) or ionizable groups (carboxylate functions) (113,114) (Fig. 17) to allow substrate recognition, as in nature, in an aqueous medium, profiting from the solvophobic effects of water (115).

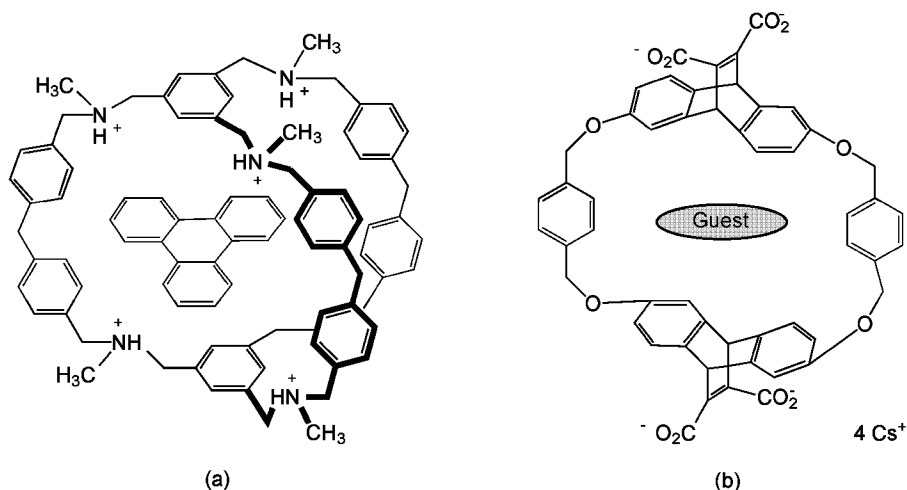


Fig. 17. Inclusion complexes of endo-hydrophobic-exopolarophilic receptors involving charges of different nature.

Multiple and Multisite, Coreceptor- and Coupled-System Substrate Recognition

Once recognition units for specific groups and individual features of a substrate have been identified, one may consider combining several of them within the same receptor. Thus far, though not carefully directed, the previous receptors in many cases, and as pointed out already possess this property of nonindividual interaction modes. More carefully directed, this leads to multiple and multisite recognition depending on the design of binding subunits which may cooperate for the simultaneous complexation of several substrates or of a multiply bound polyfunctional species to yield polynuclear complexes (homo- or heteronuclear) and mononuclear polyhapto-type complexes, respectively (116). Moreover, one may distinguish co-receptor systems for which the binding of several substrates is commutative and cascade systems, for which the substrate binding steps are noncommutative but follow a given sequence (9).

A typical example of the multiply binding type of molecular recognition of a polyfunctional species is demonstrated in Figure 18 (117). Owing to the different binding sites that are an electron deficient guanidinium group, a negatively polarized macrocyclic ring and a naphthalene π -stacking unit, the receptor shown in Figure 18a should allow simultaneous interaction with, eg, the carboxylate, the ammonium and the aromatic groups of a substrate that is an aromatic zwitterionic amino acid (tryptophan) such as illustrated in the receptor-substrate complex of Figure 18b (66,118).

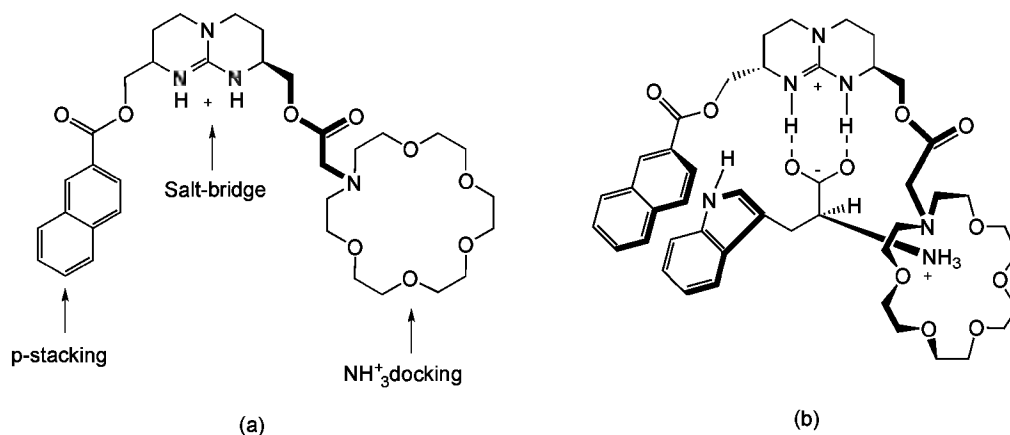


Fig. 18. (a) A bicycloguanidinium receptor; and (b) its three-point association to the zwitterionic α -amino acid tryptophan (118).

Numerous receptors comprising other multiply docking units and binding sites have been synthesized (19,35,77,117–120) including, eg, the speleands that in its frame comprise a negatively polarized macrooring and an apolar but rigid shaping component for the size selective recognition of a primary ammonium cation (121).

The macrocyclic hexamine structure of Figure 19a forms a homodinuclear cryptate with Cu(I) (122), whereas crown ether boron receptors (Fig. 19b) have been applied for the simultaneous and selective recognition of complementary cation–anion species such as potassium and fluoride (123) or ammonium and alkoxide ions (124) to yield a heterodinuclear complex (120).

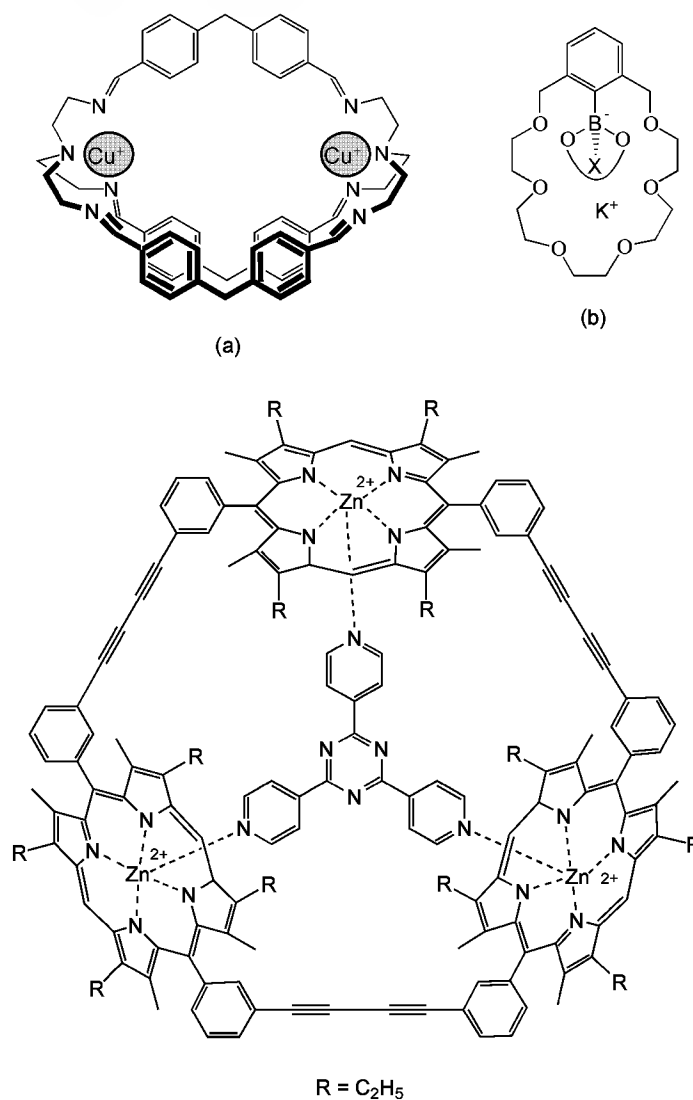


Fig. 19. Multiple and multisite substrate recognition: (a) a homo dinuclear (dicationic) and (b) a heterodinuclear (cation and anion) inclusion complex; (c) a trihapto-type inclusion complex of a triple porphyrin receptor.

Metalloreceptors based on a designed cationic inclusion complex as the specific accommodation site for the recognition of an uncharged guest molecule is another useful development (119,125). Metallomacrocycles of salen-type containing a complexed uranyl cation are most common here (126). The triple porphyrin receptor shown in Figure 19c is a more complex example of this strategy that operates in double recognition mode when forming complexes with zinc and 2,4,6-tri-4-pyridyl-*s*-triazine matching size and geometry of the tritopic metalloreceptor (127).

Receptor systems for the combined and commutative recognition of differently sized cation species (eg, K⁺ and Li⁺, Fig. 20b) have also been designed (128). They may follow the first step of the diagram in Figure 20a while recognition of the previous metalloreceptors is noncommutative with substrate binding in a given sequence typical of cascade complexation (Fig. 20a) (129). Here complexes are formed by first binding metal ions, which then serve as interaction sites for another substrate. Such is also the case for the dinuclear copper complex of Figure 20c containing a bridging imidazolato group related to the copper proteins (130).

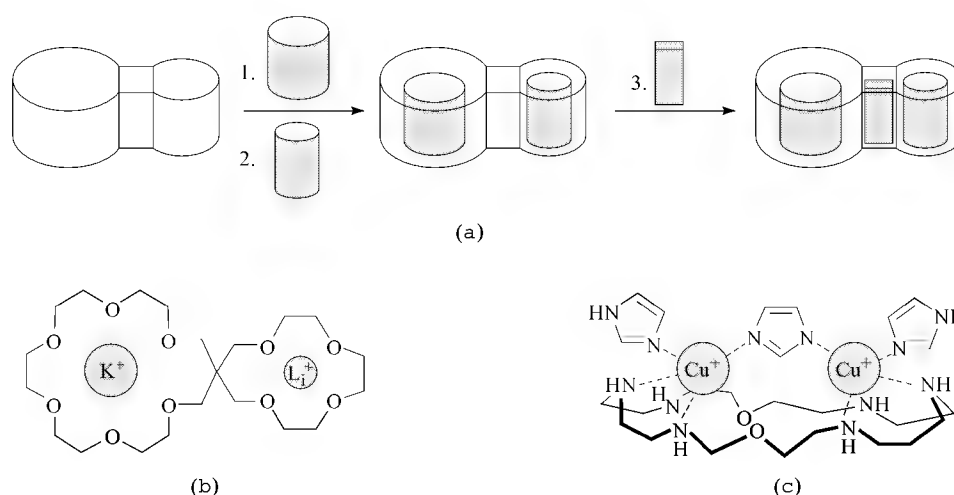


Fig. 20. (a) Schematic illustration of the formation of a cascade complex; (b) a heterodinuclear cation complex of a diloop crown receptor; and (c) a typical cascade complex.

Allosteric coupling results when occupation of a given receptor site leads to a change in the recognition and binding features of the other sites (cf Fig. 20b) making binding easier or more difficult (positive or negative cooperativity) (9,34). The allosteric effect plays a major role in the regulation of the activity of an enzyme involving conformational changes included by the binding of an effector (131). This kind of cooperativity has also been studied in synthetic receptors (132,133). An obvious approach to achieve such systems is based on the clapper-board construction schematized in Figure 21a comprising two binding sites of types A and B, respectively, inside and at the ends (34,134). Binding of M (eg, a metal ion) enhances the ability for the inclusion of L (eg, a lipophilic organic substrate). A similar case of allostery is illustrated in Figure 21b where the thymine derivative binding using hydrogen bond and π -stacking interactions is increased by a factor of 4–6 upon sodium ion uptake by the oligoethylene glycol ether part of the receptor system (135).

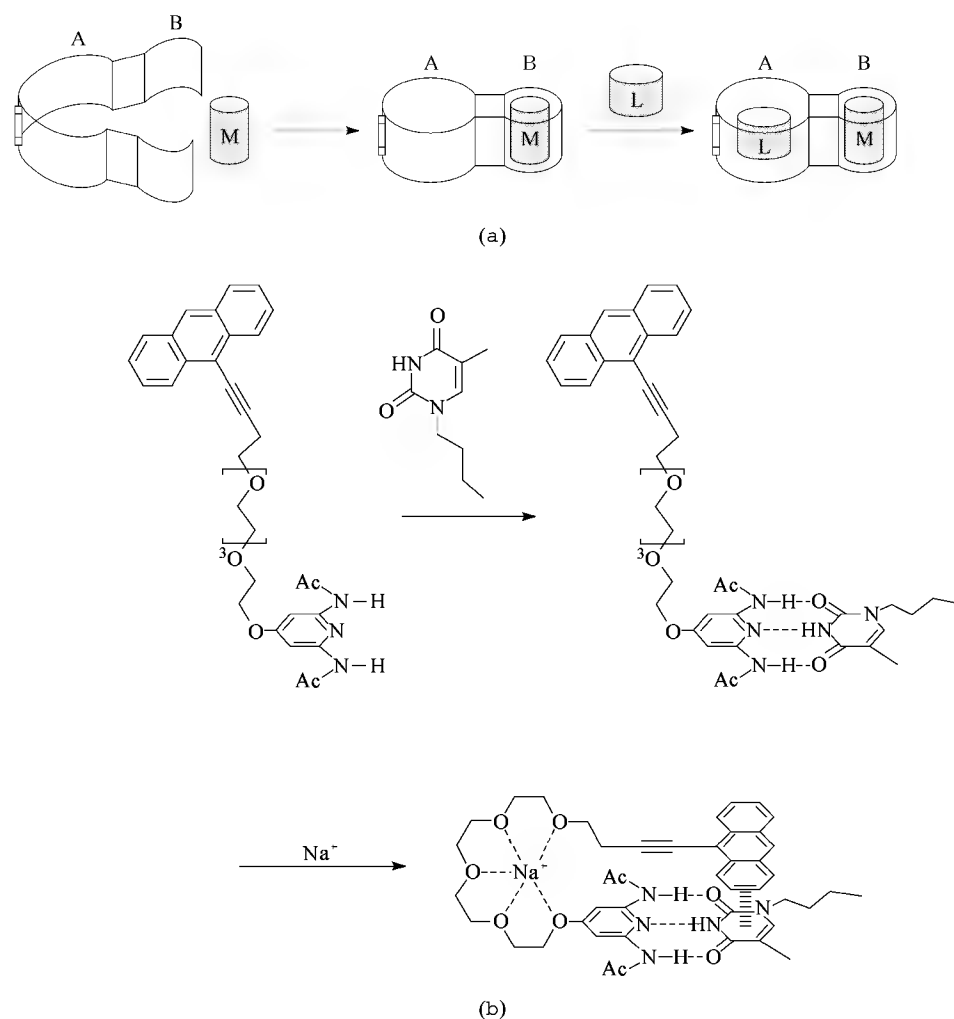


Fig. 21. (a) Diagrammatic representation of an allosteric receptor mechanism; (b) allosteric binding of a thymine derivative promoted on K^+ uptake (34).

Chiral Recognition

Enantiomers are perhaps the substrate type most difficult to distinguish. As is well known, they are stereochemical species that have exactly the same structure except for their mirror image (chirality) relationship (33) (see also CHIRAL SEPARATIONS). This causes a problem. On the other hand, chiral (enantiomer) recognition in complexation is one of the most important means by which receptor sites of biological systems such as in genes or enzymes act and regulate (2). From the principle point of view, recognition of a substrate enantiomer from racemic mixture (50:50% mixture of enantiomers) requires an enantiomeric optically resolved receptor structure in order to make possible two diastereomeric receptor–substrate complexes allowing differentiation (Fig. 22a) (136).

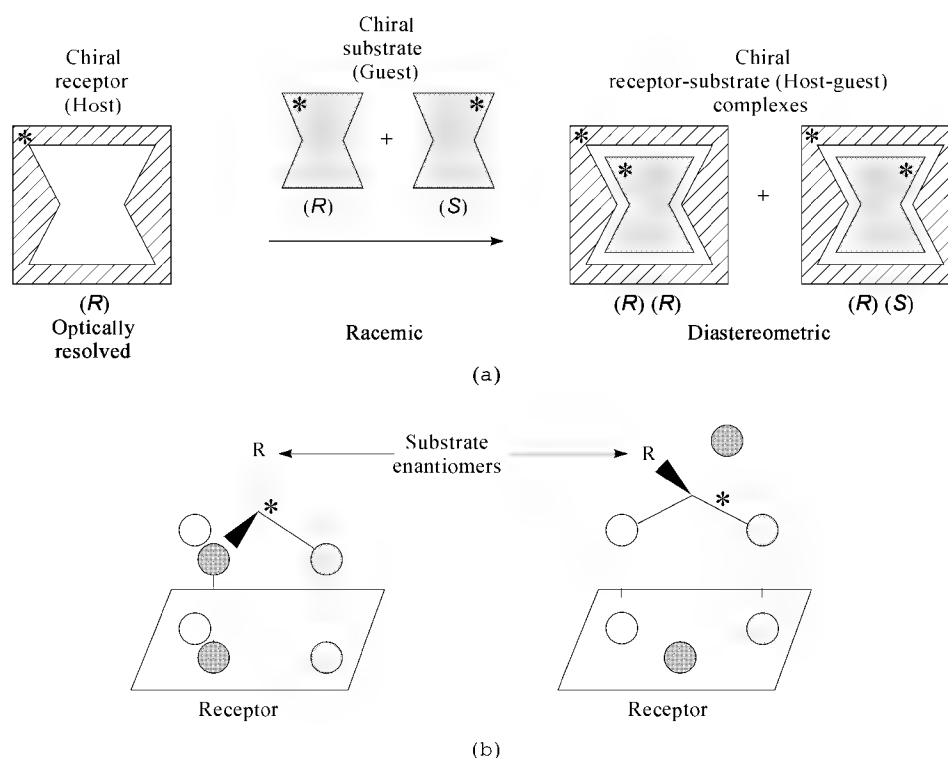


Fig. 22. Principle of chiral receptor-substrate recognition: (a) formation of diastereomeric inclusion complexes; (b) three-point interaction model.

Following this line, a great variety of optically resolved (optically active) crown compounds were prepared for the resolution of racemic cationic substrates, eg, chiral primary ammonium salts, protonated α -aminoalcohols and α -amino acid derivatives, through complexation (137–139). Among the highest enantiomer recognition properties for chiral ammonium ions were obtained with crown ethers having one or two 1,1-binaphthyl chiral barriers in the framework (Fig. 23a) (27,37,140). Others contain a spirobifluorene chiral subunit (141) or are derived from terpenoids, amino acids, and hydroxy acids that make use of the natural pool of chiral compounds (138). Typical examples for the latter classes of receptors are shown in Figure 23b, c where natural α -D-glucose (142) or tartaric acid (143) are the chiral sources. A further important family of chiral receptors derived from natural glucose are the cyclodextrins (Fig. 15a) (93–95). Moreover, most of these receptors (cf. Fig. 23a–c) are carefully designed systems in that they contain at least one C_2 axis of symmetry (dissymmetric compound type), a tactic that makes the receptors nonsided with respect to perching substrates, eg, ammonium guests (136). Beyond that C_3 symmetric receptor molecules have also been used to advantage in chiral recognition such as the cryptophane in Figure 16a (101,102) or the basket shaped chiral host presented in Figure 23d (144). Although besides chiral ammonium ions amides are the substrate class of compounds to be very efficiently resolved by the majority of receptors, particular enantiomer recognition properties including steroid hormones have also been reported (145). The cryptophane is typical of the chiral resolution of methane derivatives (eg, CHFCIBr) (146) and the basket-shaped host of Figure 23d exhibits extremely high enantioselectivity for various peptides (144).

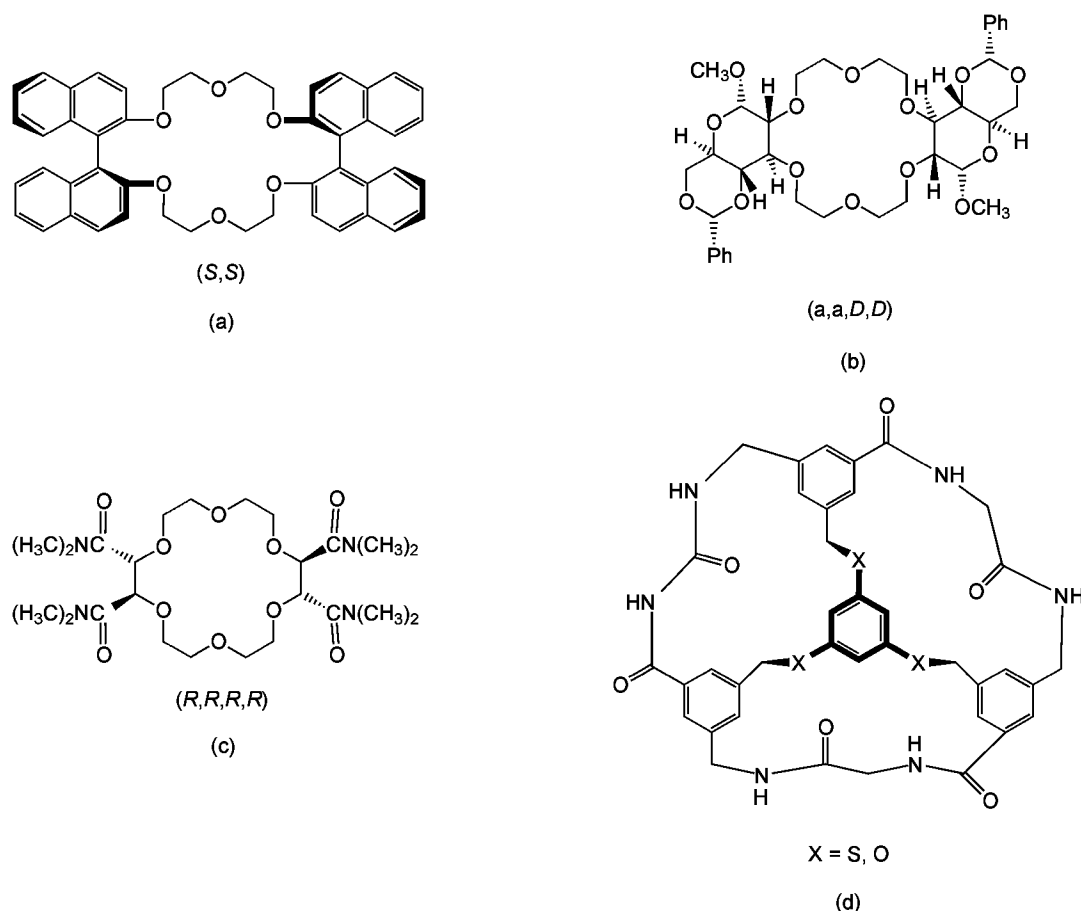


Fig. 23. Prototypical receptor molecules for chiral (enantioselective) substrate recognition.

Interpretation of these results are in keeping with the complementarity between chiral receptor and chiral substrate as sketched in Figure 22 (notice the orientation of the stars) and visualized in Figure 18b. This figure shows that the zwitterionic amino acid tryptophan of natural configuration (*S*) ideally complements the three sites of the chiral (optically resolved) receptor cleft, while the optical antipode of tryptophan (*R*-configured enantiomer) is less suited for binding (117). These facts express what is generally called the three-point interaction principle (136,147) illustrating fit or misfit of chiral receptor substrate recognition (Fig. 22b).

For more details on this topic see References 138, 139, 148; for more illustrations on the chiral fit concept see also INCLUSION COMPOUNDS.

Artificial Receptors for Particular Substrate Recognition

The recognition of substrates from many different compound classes has been discussed. Nevertheless, some particular substrates remain that are biorelevant species or play central roles as drugs. Barbiturates are such an important family of drugs and are the target for molecular recognition (19,77). According to their structure, the barbiturate moiety essentially fuses two imide groups within a six-membered ring. Thus, two diaminopyridine units correctly positioned in a macrocyclic ring should bind to all six of the accessible hydrogen bonding sites in barbiturates, as shown in Figure 24a (149). A crystal structure of a respective receptor–substrate complex has been performed that comes up to the expectations (19,77).

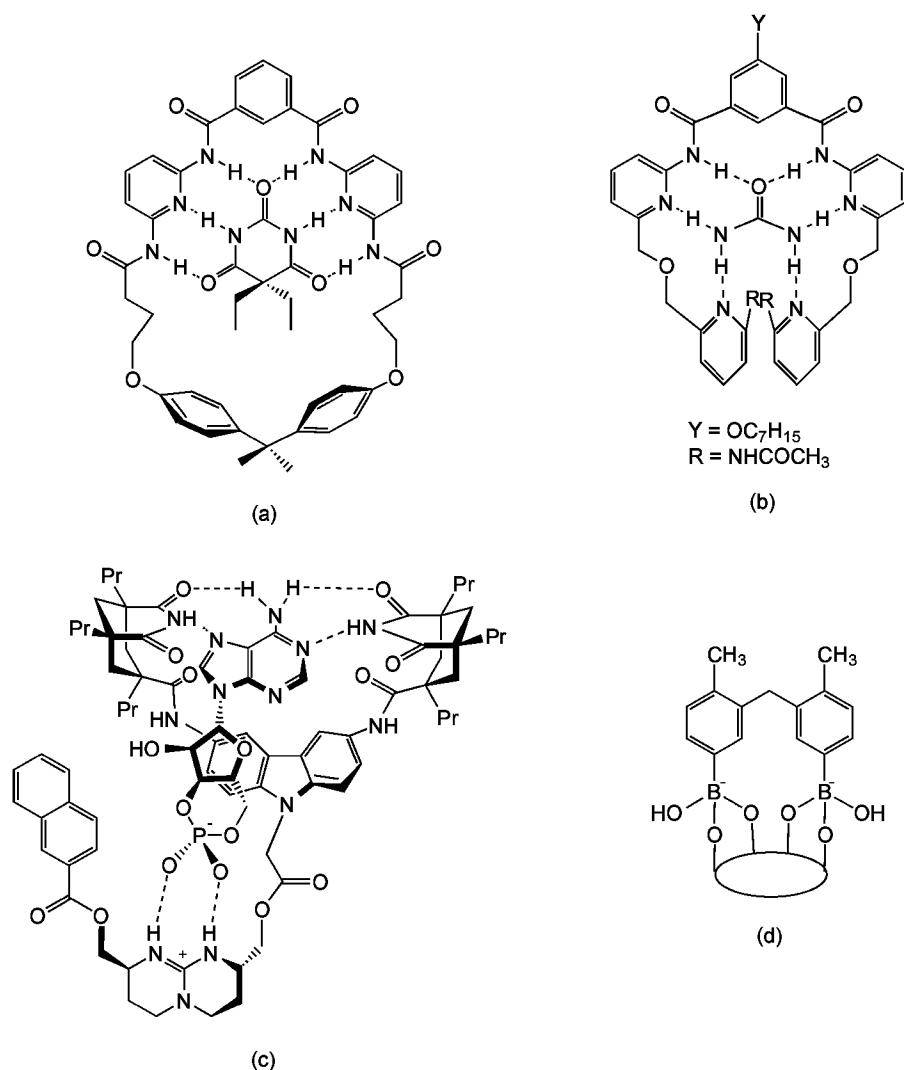


Fig. 24. Receptor substrate complexes involving particular substrate recognition.

The structural and synthetic relationships shared between barbiturates and urea, which is another substrate of high physiological interest, suggest that the above receptor strategy could be modified for the selective complexation of urea. The designed modification for urea recognition involves replacement of the H-bond donating pyridine-6-amido groups in the previous barbiturate receptor by two H-bond accepting groups that differ by 120° in alignment to the substrate. Such an arrangement of binding groups exists in the receptor molecule for urea illustrated in Figure 24b (19,77). An approach using an acyclic, but rigid hydrogen bond heteroaromatic framework such as the naphthiridine derivative presented in Figure 13d has also been developed (86), and last but not least the above metalloreceptor strategy based on a metal center (eg, UO₂) as an electrophilic binding site for the carbonyl oxygen of urea included in a macrocyclic receptor was successfully applied (119,125,126). A modification of the rigid heterocyclic cleft-type receptor mentioned above (cf Fig. 13d) has also yielded good results in the recognition of uric acid, the key product of the purine metabolism (150).

Nature's strategy for the recognition of nucleic acid bases offers an ideal example of directional and orientational dependence of simultaneous hydrogen bonding and aromatic stacking interactions to make advantage for an artificial receptor design of nucleic acid bases. A taste of it has already been given in Figures 14a (87) and 14b (88) illustrating effective recognition of an uracil and adenine derivative, respectively (19,77). In addition, receptors where two such sites are linked through a spacer element attending to the recognition of multiple nucleic acid base derivatives have also been carried out successfully (77,151).

Nucleotides, the building blocks of DNA and RNA strands, beside a polyphosphate chain and ribose or deoxyribose also contain a nucleobase. According to this importance, efforts have been directed toward the problem of phosphate recognition and the development of artificial receptors capable of distinguishing nucleotides with respect to both the nucleobase and the phosphate chain (66,118). A tricky solution of the problem is presented in Figure 24c, taking advantage of the phosphate affinities of a bicycloguanidinium cation and the well-established chelating capacity of a molecular cleft featuring two converging imide functions (152).

Boronic acid–diol covalent interactions creating five- or six-membered rings reversibly form in aqueous media, thus, providing an important tool in the recognition of saccharides (153–155). Moreover, many monosaccharides possess at least two binding sites (diol area) which differ from other monosaccharides. Based on this strategy a number of small saccharide selective receptor molecules with conformationally well-defined distance and orientation between two boronic acid functionalities have been designed. An example is given in Figure 24d (156). D-Glucose yields a relatively strong 1:1

complex at pH 11.3, whereas complexes with galactose, talose, maltose, cellobiose, and lactose are weaker. The complex with glucose is believed to involve bonds to the C1–C2 and C4–C6 diols as sketched in Figure 24d. Chiral recognition of saccharides along this line using chirally modified diboronic acid receptors have also been realized (157). A second main category of saccharide receptors are typical of bowl-shaped molecules belonging to the resorcinarenes (158) (see INCLUSION COMPOUNDS) and other more recent examples (153). What is more, there are good expectations for the development of artificial adrenalin (159) and peptide receptors (160).

However, all the receptors hitherto discussed are monomolecular species which possess a monomolecular cavity, pocket, cleft, groove or combination of it including the recognition sites to yield a molecular receptor–substrate complex. They can be assembled and preserved in solution although there are dependences (see below). By way of contrast, molecular recognition demonstrated in the following comes from multimolecular assembly and organization of a nonsolution phase such as polymer materials and crystals.

Molecular Recognition in Polymers and Solids

If a polymer is prepared in the presence of molecules, the "print molecules" which are extracted after polymerization, the remaining polymer may contain cavities, prints, or footprints that can recognize the print molecule (161). Actually, the cast relates to the matrix molecule like lock and key of Emil Fischer's long-known principle (5) (see Fig. 1). A scheme of this clever imprinting technique is illustrated in Figure 25a, and a relevant example of imprinted polymer is shown in Figure 25b. The template monomer of Figure 25b is a compound in which two polymerizable moieties of vinylphenylboronic acid have been bound to phenyl α -D-manopyranoside as the print molecule. A 1:9 ratio of the monomer and ethylene dimethacrylate were copolymerized in an inert solvent to yield a macroporous polymer of which the template is split off providing chiral cavities each bearing a pair of boronic acid groups. Polymers of this type have excellent ability for recognition of the template molecule, also in the given enantiomer configuration when subjected to its racemate.

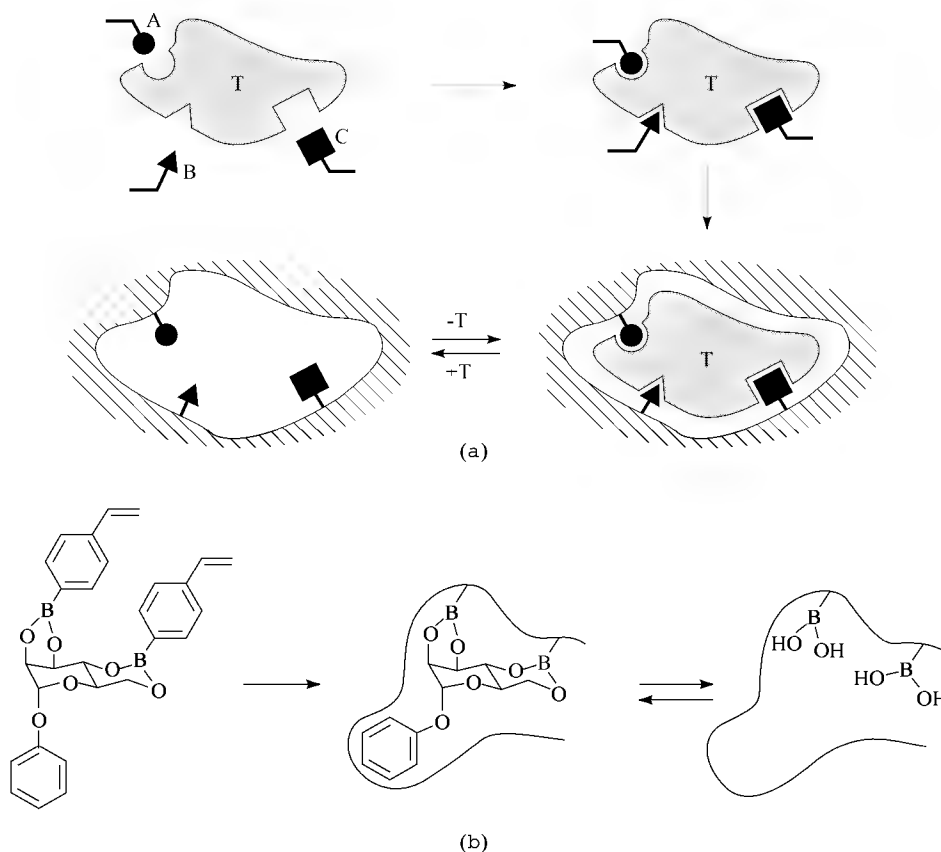


Fig. 25. Schematic representation of imprinting: (a) cross-linking polymerization in the presence of a template (T) to obtain cavities of specific shape and a defined spatial arrangement of functional groups (binding sites, A–C); (b) cross-linked polymer prepared from the template monomer and ethylene dimethacrylate with and without the templates (161).

By analogy, a great many of other functionalized styrenes, including carboxylic acids, amino acids, Schiff bases, or specific compounds, eg, L-DOPA, have successfully been applied as print templates. Moreover, it has also been shown that silica gel can be imprinted with similar templates, and that the resulting gel has specific recognition sites determined by the print molecule (162–164).

In a sense, molecular recognition using hollow organic crystals, in particular clathrate structures (165), is similar although the interactions forming the framework are noncovalent but weak interactions (16). For the same reason typical multinuclear crystalline inclusion compounds and clathrates are not stable in solution, but decompose unlike the monomolecular inclusions and more stable receptor–substrate complexes (165). Nevertheless molecular recognition behavior of crystalline inclusion compounds is both various dependent on the structure that can be cavity-, layer- or channel-type (167), and in many cases highly selective including chiroselectivity (168).

A packing motif giving a general idea of such an efficient chiral recognition machinery that uses the crystalline inclusion phenomenon is illustrated in Figure 26 (169). The optically active receptor molecule (Fig. 26a), a bulky derivative of lactic acid, chiroselectively yields a crystalline inclusion compound with (R)-configured 3-methylcyclohexanone (Fig. 26b), refusing the steric mirror image (*S*-configuration) of the substrate, whereas the correctly configured substrate (*R*) ideally matches the intermolecular lattice space (Fig. 26c). Along this line a number of bulky crystalline hosts have been designed capable of chiral substrate recognition, eg, of alcohols, phenyloxirane, sulfoxides and lactones, to say nothing of more simple constitutional isomer recognition and of the recognition of other chemically different species (165,167,170). However, it would mean doubling of information to go into details here since this particular topic is extensively covered under the subjects extra molecular cavity and lattice type inclusion compounds (see INCLUSION COMPOUNDS).

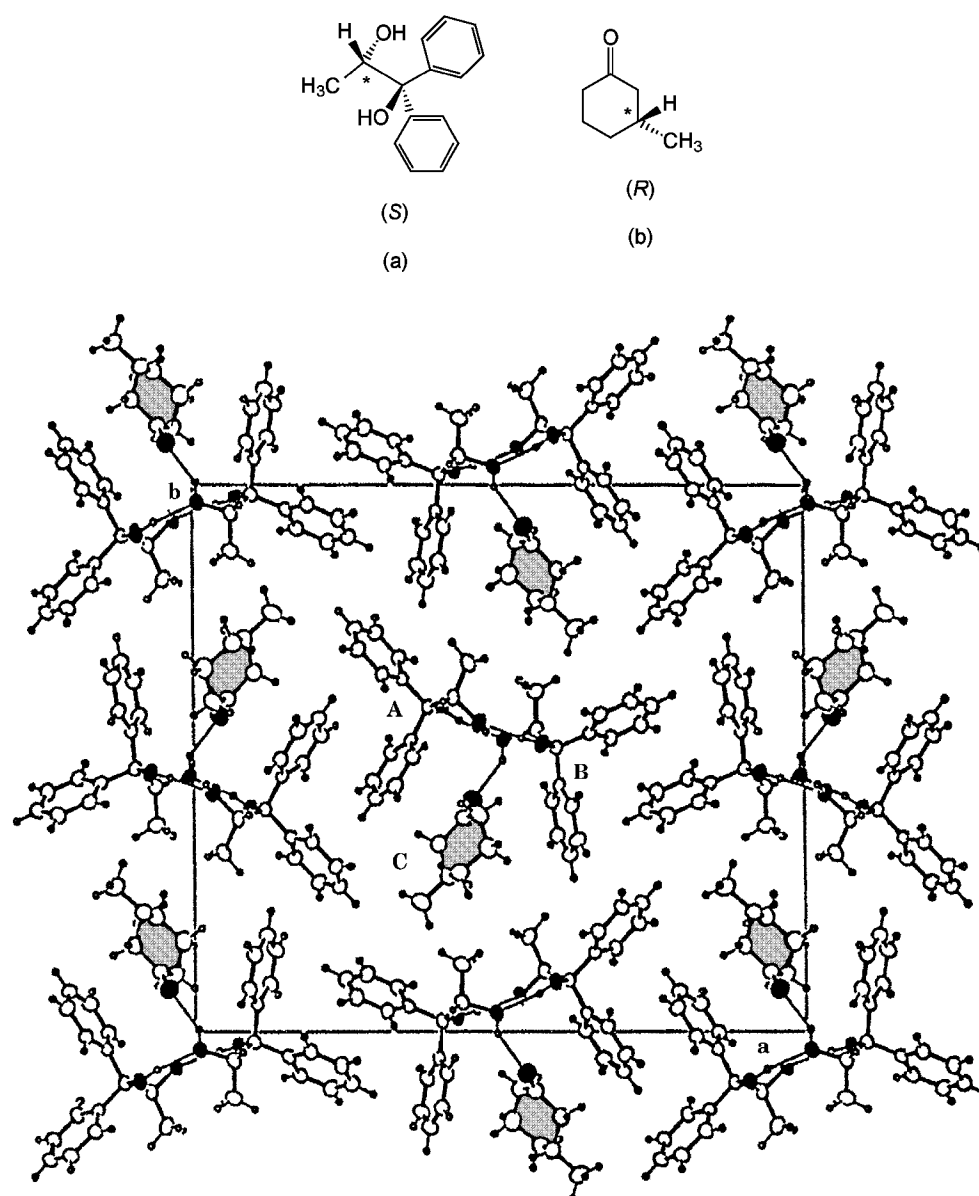


Fig. 26. Clathrate receptor chemistry: (a) a chiroselective crystalline host compound (clathrand); (b) a typical guest molecule to be included in the specified configuration; and (c) the crystal structure of the respective clathrate (A and B denote host and C the guest species) (169).

Microporous inorganic materials dominated historically by the zeolites and aluminosilicates, and the great variety of more recent nonoxide and coordination framework materials should also be mentioned here (171–174) but not discussed in detail. This type of molecular recognition is usually known as molecular sieving.

Molecular Recognition at Interfaces and Surface Monolayers

There are three advantages to study molecular recognition on surfaces and interfaces (monolayers, films, membranes or solids) (175): (1) rigid receptor sites can be designed; (2) the synthetic chemistry may be simplified; (3) the surface can be attached to transducers which makes analysis easier and may transform the molecular recognition interface to a chemical sensor. And, which is also a typical fact, this kind of molecular recognition involves outside directed interaction sites, ie, exo-receptor function (9) (see Fig. 5b).

To begin with, molecular recognition of crystal interfaces make possible the control of crystal growth processes in that suitably designed auxiliary molecules act as promoters or inhibitors of crystal nucleation inducing, for instance, the resolution of enantiomers or the crystallization of desired polymorphs and crystal habits (176). As an example (Fig. 27), crystals of achiral glycine, due to their enantiopolar arrangement, may differentiate between (R)- and (S)-amino acids when being used as additives (177). Thus an α -amino acid additive such as alanine of configuration (R) capable of replacing a glycine molecule will block crystal growth only at one of the two enantiopolar crystal faces, the (*pro*-R)-face [(010)-face in Fig. 27b]. An (S)-amino acid practises the same effect on the (*pro*-S)-face of the glycine crystal [(010)-face], and [(0 $\bar{1}$ 0)-face], and addition of racemic (R,S)-amino acid suppresses growth at both enantiopolar faces (Fig. 27c). In consequence, the crystal of glycine changes its habit from a symmetric bipyramid to either two asymmetric mirror-image pyramids or a platelet making engineered crystal habits possible (178) (Fig. 27). These impurity crystals having large and hydrophobic planes orientedly swim at the air–water surface which can be applied toward the resolution of further amino acids and for the direct and relative assignment of the absolute configuration of chiral molecules and crystals using stereochemical correlations (179).

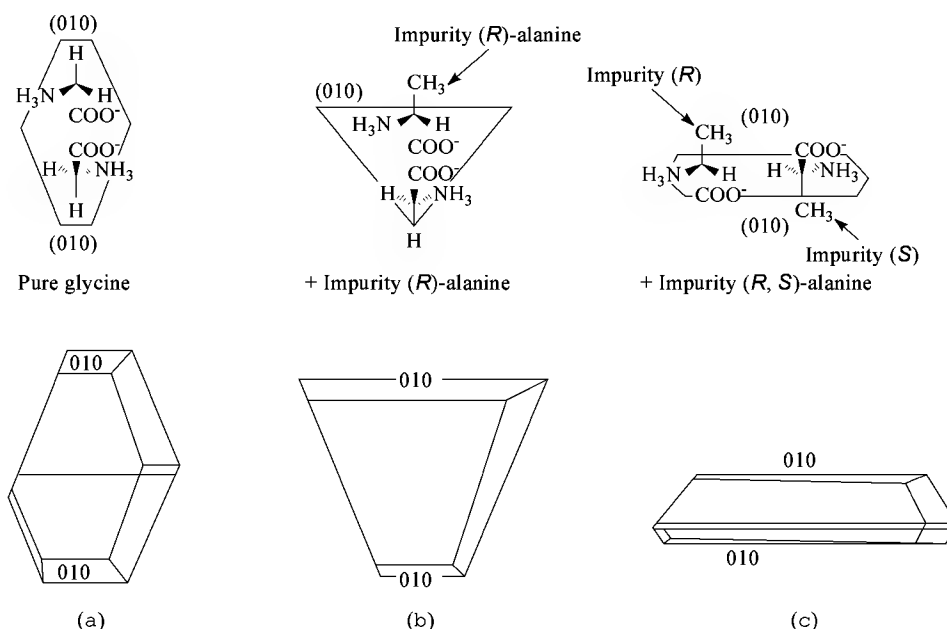


Fig. 27. Recognition at crystal interfaces and its role in the engineering of crystal morphology and configurational assignment of molecules (176,177).

Following another direction, it has previously been shown that alkanethiols spontaneously adsorb to Au from dilute solutions of ethanol and other nonaqueous solvents, and that the resulting self-assembling monolayers (SAMs) assume a close-packed overlayer structure on Au (111) and other textured Au surfaces, being quite robust in aqueous solutions and vapor-phase ambients (180). This mode of self-assembly chemistry has been used to synthesize monolayer assemblies that function as molecular recognition interfaces based on the presence of recognizer end groups (181). Thus one-component SAMs formed of *n*-alkanethiols having extra carboxylic acid functionalized end groups specifically adsorb vapor-phase acid-terminated molecules via H-bonding (Fig. 28) or vapor-phase amine-terminated molecules via proton-transfer interaction, exhibiting chemical complementorship (182). It has also been demonstrated that two-component SAMs, which consist of inert *n*-alkanethiol framework molecules and defect inducing template molecules, can discriminate between solution-phase probe molecules based on their geometrical properties, similar to the imprinting technique discussed before (see Fig. 25a) but on the two-dimensional level only (182). In order to create analogous molecular recognition property on an oxide surface, a silica overlayer was prepared on tin oxide by the chemical vapor deposition of silicon alkoxide using preadsorbed benzoate anions as template molecules, which were then removed to yield vacancies in the overlayer capable of size recognition of the probe molecules (183).

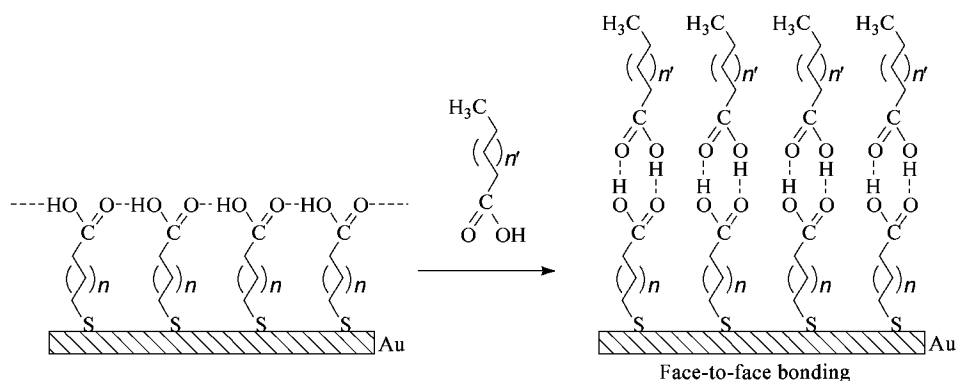


Fig. 28. Self-assembly monolayer to function as molecular recognition interface (182).

In an obvious example of molecular recognition at the air–water interface, the receptor consists of an organized monolayer formed from amphiphiles that have complementary binding sites (184). For example, the double-chain triazine amphiphile illustrated in Figure 29 (185) was employed for the formation of a monolayer receptor at the air–water interface which specifically interacts with barbituric acid dissolved in the water subphase, creating a supramolecular strand, in close analogy to solid-state and solution structures formed of the hydrogen-bonded components (186,187). The stabilization conferred on the monolayer by its networking with barbituric acid made its imaging by atomic force microscopy (afm) possible, while the noncomplexed monolayer is destroyed by the scanning tip of the afm (185). Vice versa, amphiphilic derivatives of barbituric acid form a monolayer receptor at the air–water interface interacting with complementary substrates such as 1,4,6-triaminopyridine, melamine and urea (188). The recognition reaction that occurred was investigated by measuring the pressure–surface area isotherms and by using uv-visible spectroscopy. For this latter reason some of the recognizer amphiphiles contain the azobenzene chromophore in its framework (189). Various monolayer self-assembled systems developed primarily for the recognition of nucleobases, dipeptides and sugars at the air–water interface including amphiphilic recorcinarene receptors have also been studied (184). Furthermore, monolayer receptors originating from long-chain derivatives of Kemp's acid (see Fig. 13c) were employed for the molecular recognition of amino acids and various nitrogen aromatic compounds (190).

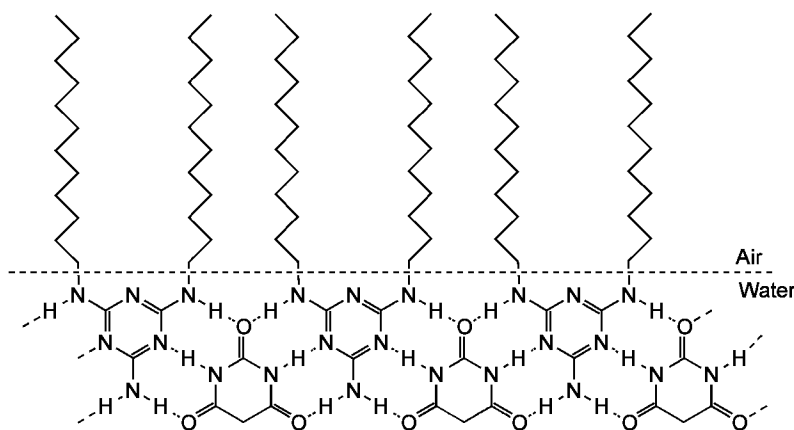


Fig. 29. Molecular recognition of an organized assembly at the air–water interface (184).

In another example, involving vesicles as organizing entities, molecular recognition occurred between mixed vesicles bearing recognizable moieties. Interaction occurs between the mixed vesicles, which leads to the formation of larger aggregates attributable to the interaction of their recognizable moieties located at the external interface of the vesicles (191).

For the investigation of molecular recognition in micelles, adenine derivatives and positively charged (thyminyllalkyl)ammonium salts such as shown in Figure 30 were prepared, which were solubilized in sodium dodecyl sulfate (SDS) solutions. Nmr studies have shown that binding occurs in a 1:1 molar ratio in the interior of the micelles as illustrated in Figure 30 (192).

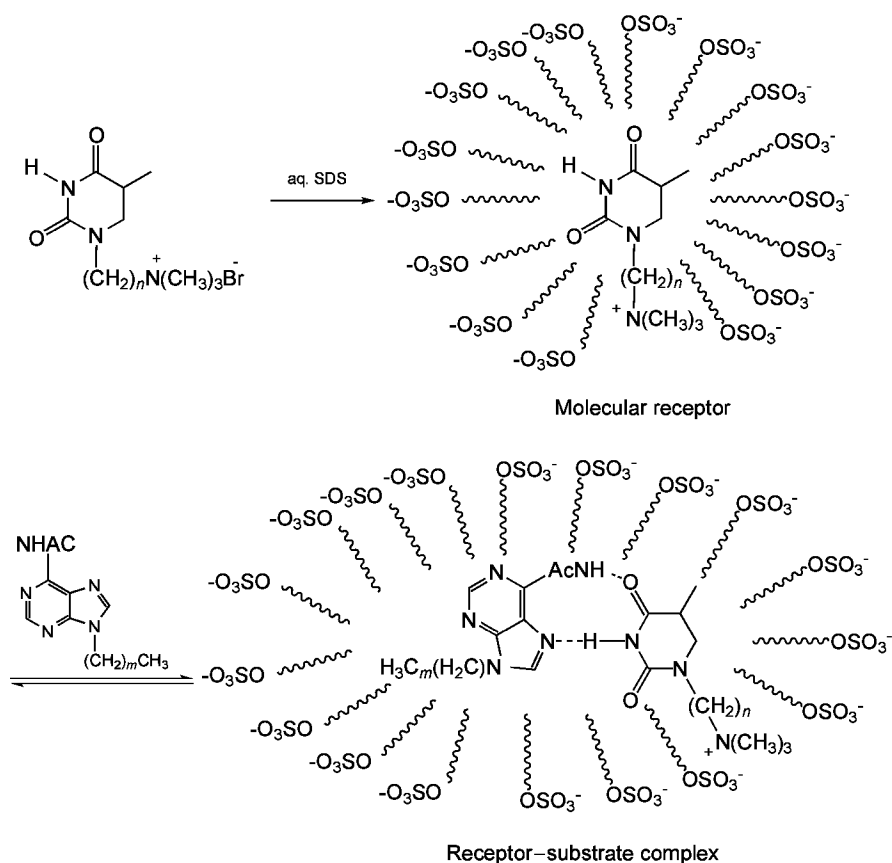


Fig. 30. Molecular recognition in micellar systems (184).

In summary, one may say that the air–water inface is much larger, smoother, and more organized than the interfaces of aqueous micelles and bilayers, which in turn are larger and more organized than those of molecular pairs in bulk water. Thus the binding affinity in general is increased by a factor of 10^2 – 10^4 each time one proceeds from bulk water to microscopic interfaces of micelles or bilayers and finally to air–water macroscopic interfaces having obvious bearing on the recognition property (184).

Self-Recognition

This mode of molecular recognition, on principle, is defined as the recognition of like from unlike or self from unself-molecules, embodied in the spontaneous selection and preferential assembly of like components in a mixture (9).

So far this article has been concerned with interactions among chemically different species which is true for most of the chemical recognition processes including supramolecular and biomolecular processes. With crystals it is usually the other way around. Although some crystals, co-crystals, crystalline complexes, and crystalline inclusion compounds (see above, and INCLUSION COMPOUNDS), are built from more than one kind of molecule and are exceptions, most crystals are built from identical (or enantiomeric) copies of the same molecule. Thus, a usual one-component crystal is a macro-supramolecular assembly (193) where one should more properly speak of molecular self-recognition. This is not a fact contradictory to the basic principles of molecular recognition, discussed at the beginning, since naturally the case might occur in which the two complementary structures happen to be identical in dealing with a self-complementary relationship. So, even when all the molecules are identical (or enantiomeric), an acceptor part of one molecule can interact with a donor part of a second, and the acceptor part of the second can interact in exactly the same manner with the donor part of a third, and so on, giving rise to periodicity of the crystal and to the limited number of space groups used in molecular crystals (194). For instance, it is very uncommon for molecules in a crystal structure to be related by rotation axis or mirror planes, because identical parts of molecules avoid one another,

except for molecular sites having a so-called self-complementary donor–acceptor group (195). Self-complementary groups such as the carboxylic acid, the amide, the urea function or its combinations form finite, one-dimensional tape, two-dimensional layer, or three-dimensional motifs of organic molecules mostly obtained from hydrogen bonding. Representative examples are given in Figure 31 (196).

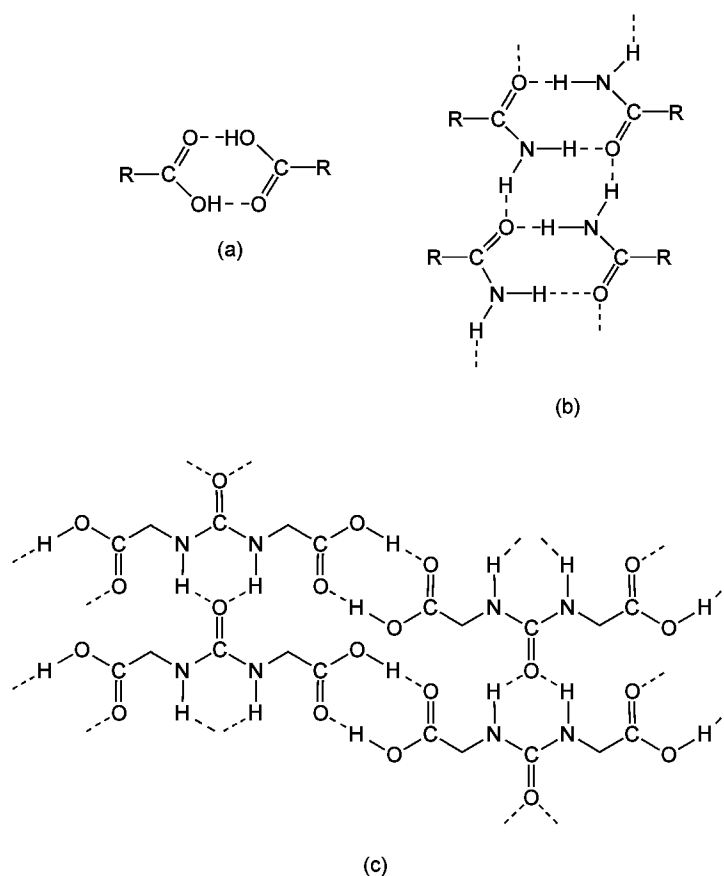


Fig. 31. Supramolecular (hydrogen-bonded) motifs of self-complementary molecules (196).

In solution, highly ordered structures created via self-recognition and self-assembly of a programmed H-bonding molecular component are also possible (197) such as the hexameric pseudo supermacrocycle of a designed DNA base hybrid shown in Figure 32 (198), to say nothing of the self-assembly of organic Langmuir and Langmuir-Blodgett films (199) where self-recognition at the air–water interface is of vital importance as well (see Fig. 29).

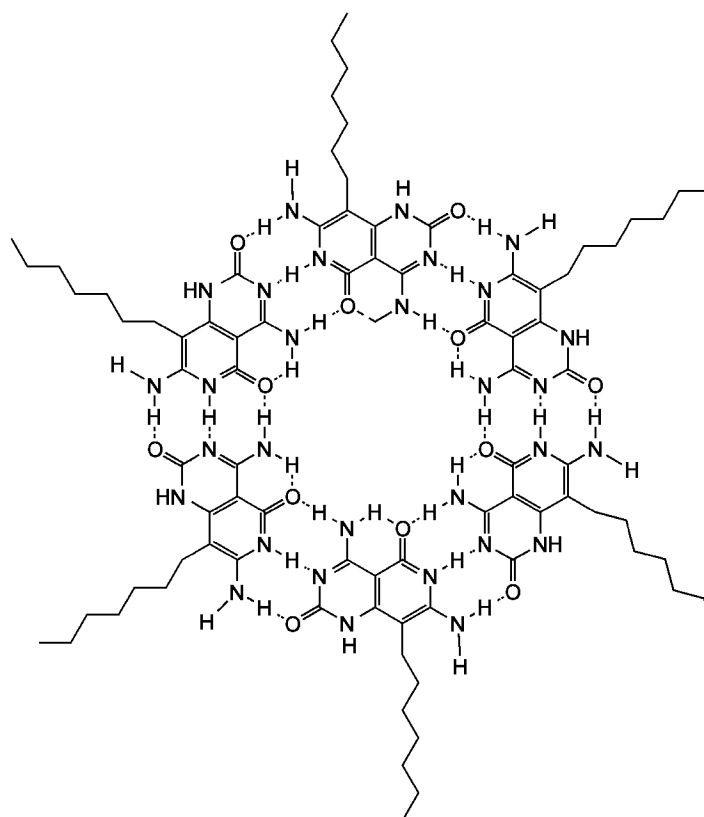


Fig. 32. Hexameric self-assembled supermolecule involving self-recognition (198).

With respect to inorganic self-recognition and self-assembly this would involve preferential binding of like metal ions by like ligands in a mixture of ligands and ions (9). Indeed, selective formation of double-helicates was obtained from mixtures of oligo-bipyridine strands in the presence of suitable metal ions, eg, Cu^+ , without significant crossover, ie, the desired helicates are generated with self-recognition (Fig. 33) (200). Similar self-recognized

triple-helicates have also been obtained from bis-catechols and Ga(III) (201).

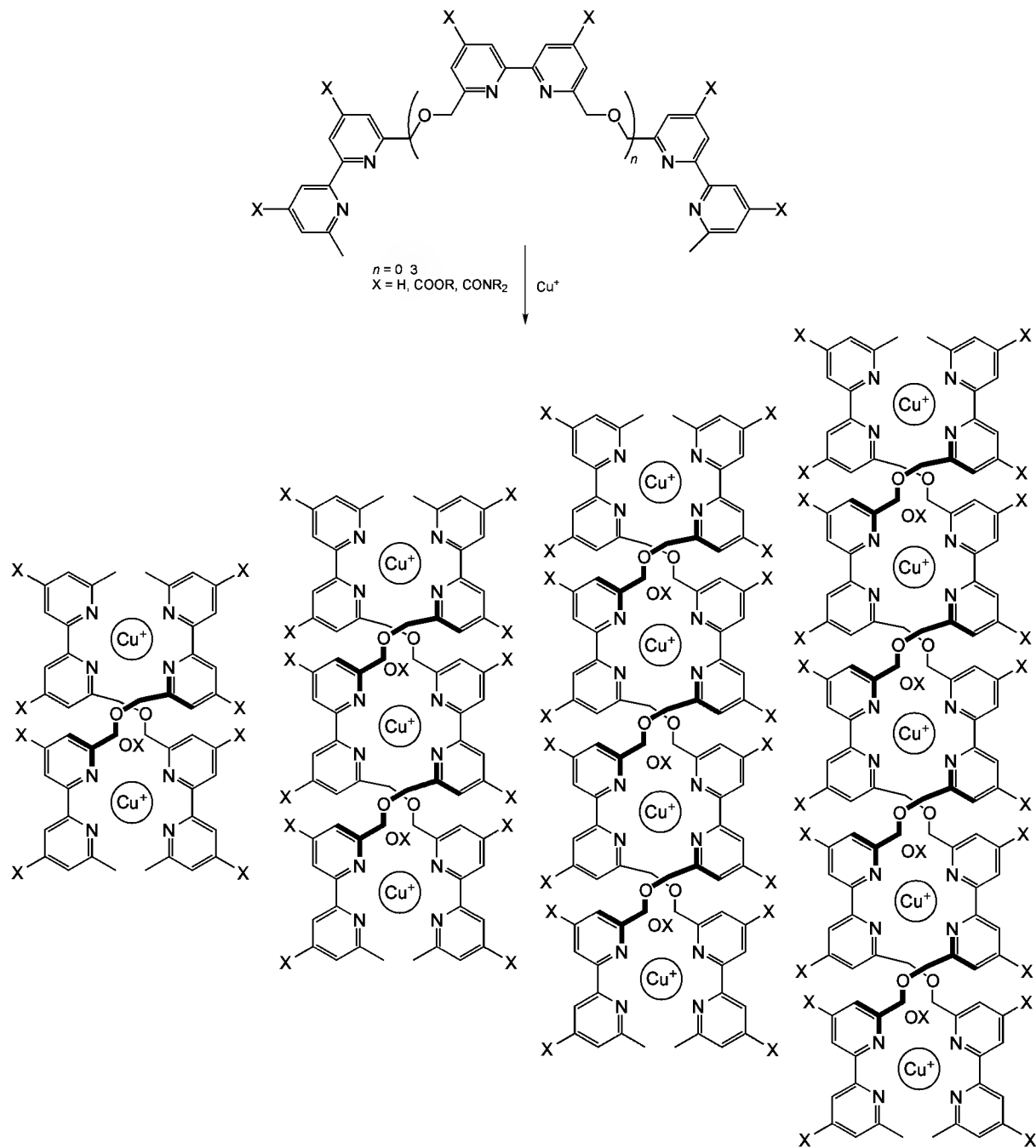


Fig. 33. Self-recognition in the self-assembly of double helices (9,200).

A particular point of interest included in these helical complexes concerns the chirality. The helicates obtained from the achiral strands are a racemic mixture of left- and right-handed double helices (Fig. 34) (202). This special mode of recognition where homochiral supramolecular entities, as a consequence of homochiral self-recognition, result from racemic components is known as optical self-resolution (203). It appears in certain cases from racemic solutions or melts (spontaneous resolution) and is often quoted as one of the possible sources of optical resolution in the biological world. On the other hand, the more commonly found process of heterochiral self-recognition gives rise to a racemic supramolecular assembly of enantio pairs (204).

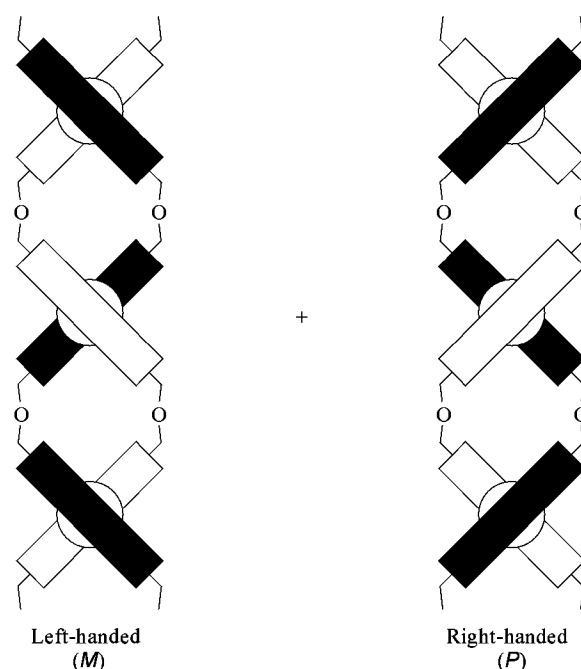


Fig. 34. Enantiomeric double-stranded helices corresponding to Figure 33 (9,202).

Conclusion

Although molecular recognition, the programmed molecular interaction or binding with a purpose, is difficult to put into numerical values, its fundamental importance in natural sciences with special points of emphasis in molecular biology (1–3) and supramolecular chemistry (8–12) is clear now. Of course, one may say that molecular recognition as itself is old hat, considering the early lock-and-key principle of Emil Fischer (5) mentioned at the beginning. Nevertheless, molecular recognition has also proved the main portal to enter into the new fascinating technological developments that supramolecular sciences have made available and may further open in the future (205).

BIBLIOGRAPHY

1. M. I. Page, *The Chemistry of Enzyme Action*, Elsevier, Amsterdam, 1984.
2. E. C. Hulme, *Receptor Biochemistry*, Oxford University Press, New York, 1990.
3. S. M. Roberts, ed., *Molecular Recognition—Chemical and Biochemical Problems*, The Royal Society of Chemistry, Cambridge, 1989.
4. J.-P. Behr, ed., *The Lock and Key Principle, Perspectives in Supramolecular Chemistry*, Vol. 1, Wiley, Chichester, 1994.
5. E. Fischer, *Ber. Dtsch. Chem. Ges.* **27**, 2985 (1894); see also F. W. Lichtenthaler, *Angew. Chem.* **106**, 2456 (1994); *Angew. Chem. Int. Ed. Engl.* **33**, 2364 (1994).
6. P. Ehdich, *Studies on Immunity*, Wiley, New York, 1906.
7. A. Werner, *Z. Anorg. Chem.* **3**, 267 (1893).
8. F. Vögtle, *Supramolecular Chemistry—An Introduction*, Wiley, Chichester, 1993.
9. J.-M. Lehn, *Supramolecular Chemistry*, VCH, Weinheim, 1995.
10. E. Weber, ed., *Supramolecular Chemistry I—Directed Synthesis and Molecular Recognition*, *Top. Curr. Chem.*, Vol. 165, Springer, Berlin-Heidelberg, 1993.
11. E. Weber, ed., *Supramolecular Chemistry II—Host Design and Molecular Recognition*, *Top. Curr. Chem.*, Vol. 175, Springer, Berlin-Heidelberg, 1995.
12. J. L. Atwood, J. E. Davies, D. D. MacNicol and F. Vögtle, eds., *Comprehensive Supramolecular Chemistry*, Vols. 1–10, Elsevier, Oxford, 1996.
13. F. Vögtle and E. Weber, eds., *Host-Guest Complex Chemistry—Macrocycles*, Springer, Berlin-Heidelberg, 1985.
14. J.-M. Lehn, *Struct. Bonding* **16**, 1 (1973).
15. H.-J. Schneider, T. Blatter, U. Cuber, R. Juneja, T. Schiestel, U. Schneider, I. Theis, and P. Zimmermann, in H.-J. Schneider and H. Dürr, eds., *Frontiers in Supramolecular Chemistry and Photochemistry*, VCH, Weinheim, 1991, p. 29.
16. M. Mascal, *Contempor. Org. Syn.*, 31 (1994); see also G. R. Desiraju, *Angew. Chem.* **107**, 2541 (1995); *Angew. Chem. Int. Ed. Engl.* **34**, 2311 (1995).
17. H.-J. Schneider, *Chem. Soc. Rev.*, 227 (1994).
18. D. E. Koshland, Jr., *Angew. Chem.* **106**, 2468 (1994); *Angew. Chem. Int. Ed. Engl.* **33**, 2375 (1994).
19. D. Hamilton, in H. Dugas, ed., *Bioorganic Chemistry Frontiers*, Vol. 2, Springer, Berlin-Heidelberg, 1991, p. 115.
20. D. J. Cram, *Angew. Chem.* **100**, 1041 (1988); *Angew. Chem. Int. Ed. Engl.* **27**, 1009 (1988).
21. E. Weber in S. Patai and Z. Rappoport, eds., *Crown Ethers and Analogs*, Wiley, Chichester, 1989, p. 305.
22. F. Vögtle and E. Weber, *Angew. Chem.* **91**, 813 (1979); *Angew. Chem. Int. Ed. Engl.* **18**, 753 (1979).
23. G. W. Gokel and O. Murillo, in Ref. 12, Vol. 1, p. 1.
24. G. W. Gokel, *Crown Ethers and Cryptands, Monographs in Supramolecular Chemistry*, Vol. 3, The Royal Society of Chemistry, Cambridge, 1991.
25. J. M. Lehn, *Acc. Chem. Res.* **11**, 49 (1978).
26. B. Dietrich, P. Viout, and J. M. Lehn, *Macrocyclic Chemistry*, VCH, Weinheim, 1993.
27. D. J. Cram and K. N. Trueblood, in F. Vögtle, ed., *Host Guest Complex Chemistry I, Top. Curr. Chem.*, Vol. 98, Springer, Berlin-Heidelberg, 1981, p. 43; see also D. J. Cram and K. N. Trueblood, in Ref. 13, p. 125.
28. D. J. Cram, *Angew. Chem.* **98**, 1941 (1986); *Angew. Chem. Int. Ed. Engl.* **25**, 1039 (1986).
29. J.-M. Lehn, *Angew. Chem.* **100**, 91 (1988); *Angew. Chem. Int. Ed. Engl.* **27**, 89 (1988).
30. E. Weber and F. Vögtle, in Ref. 12, Vol. 2, p. 1.
31. D. J. Cram, *From Design to Discovery*, American Chemical Society, Washington, D.C., 1990.
32. A. G. Amit, R. A. Marínzsa, S. E. V. Phillips, and R. J. Poljak, *Science* **233**, 747 (1986).
33. S. Hauptmann and G. Mann, *Stereochemie*, HTB Spektrum Akademischer Verlag, Heidelberg, 1996, p. 291.
34. H.-J. Schneider and A. K. Mohammad-Ali, in Ref. 12, Vol. 2, p. 69.
35. A. R. van Doorn, W. Verboom and D. N. Reinhoudt, in G. W. Gokel, ed., *Advances in Supramolecular Chemistry*, Vol. 3, JAI Press, Greenwich, 1993, p. 159.
36. B. Künnig, *J. Prakt. Chem.* **337**, 339 (1995).

37. C. J. Pedersen, *Angew. Chem.* **100**, 1053 (1988); *Angew. Chem. Int. Ed. Engl.* **27**, 1021 (1988).
38. E. Weber and F. Vögtle, in Ref. 13, p. 1.
39. E. Weber and H.-P. Josel, *J. Incl. Phenom.* **1**, 79 (1983).
40. R. Hilgenfeld and W. Saenger, in F. Vögtle, ed., *Host-Guest Complex: Chemistry II, Top. Curr. Chem.* Vol. 101, Springer, Berlin-Heidelberg, 1982, p. 1; see also, R. Hilgenfeld and W. Saenger, in Ref. 13, p. 43.
41. G. W. Gokel and O. F. Schall, in Ref. 12, Vol. 1, p. 97.
42. M. Dobler, *Ionophores and Their Structures*, Wiley, New York, 1981.
43. M. R. Truter, *Struct. Bonding* **16**, 71 (1973).
44. Y. Inoue and G. W. Gokel, eds., *Cation Binding by Macrocycles*, Marcel Dekker, New York, 1990.
45. T. Suzuki, K. Nakashima, and S. Shinkai, *Chem. Lett.*, 699 (1994); see also T. Anderson, K. Nilson, M. Sundahl, G. Wetsman, and O. Wennerström, *J. Chem. Soc., Chem. Commun.*, 604 (1992).
46. J.-M. Lehn, *Pure Appl. Chem.* **49**, 857 (1977).
47. E. Graf and J.-M. Lehn, *J. Am. Chem. Soc.* **97**, 5022 (1975).
48. E. Graf and J.-M. Lehn, *Helv. Chim. Acta* **64**, 1040 (1981).
49. E. Graf, J.-P. Kintzinger, J.-M. Lehn, and J. LeMoigne, *J. Am. Chem. Soc.* **104**, 1672 (1982).
50. J.-M. Lehn and P. Vieding, *Tetrahedron Lett.* **21**, 1323 (1980).
51. J. C. Metcalfe, J. F. Stoddart and G. Jones, *J. Am. Chem. Soc.* **99**, 8317 (1977).
52. J. Krane and O. Aune, *Acta Chem. Scand.* **B 34**, 397 (1980).
53. J.-M. Lehn, P. Vieding, and R. C. Hayward, *J. Chem. Soc., Chem. Commun.*, 296 (1979).
54. J. W. H. M. Uiterwijk, S. Harkema, J. Gevers, and D. N. Reinhoudt, *J. Chem. Soc., Chem. Commun.*, 220 (1982).
55. I. O. Sutherland, in G. W. Gokel, ed., *Advances in Supramolecular Chemistry*, Vol. 1, JAI Press, Greenwich, 1990, p. 65.
56. I. O. Sutherland, *J. Incl. Phenom.* **7**, 213 (1989).
57. C. Pascard, C. Riche, M. Cesario, F. Kotzyba-Hibert, and J.-M. Lehn, *J. Chem. Soc., Chem. Commun.*, 557, (1982).
58. J. S. Bradshaw, K. E. Krakowiak, and R. M. Izatt, *Azo-Crown Macrocycles*, Wiley, New York, 1993.
59. J. S. Bradshaw and J. Y. K. Hui, *J. Heterocycl. Chem.* **11**, 649 (1974).
60. S. R. Cooper, *Acc. Chem. Res.* **21**, 141, (1988).
61. S. R. Cooper, ed., *Crown Compounds: Toward Future Applications*, VCH, Weinheim, 1992.
62. T. L. Ho, *Hard and Soft Acids and Bases Principle in Organic Chemistry*, Academic Press, New York, 1977.
63. K. N. Raymond, *Coord. Chem. Rev.* **105**, 135 (1990).
64. F. Ebmeyer and F. Vögtle, in H. Dugas, ed., *Bioorganic Chemistry Frontiers*, Vol. 1, Springer, Berlin-Heidelberg, 1990, p. 143.
65. K. Gloe, H. Stephan, O. Heitzsch, H. Bukowsky, E. Uhlemann, R. Pollex, and E. Weber, *J. Chem. Soc., Chem. Commun.*, 1955 (1994).
66. C. Seel, A. Galón, and J. de Mendoza, in Ref. 11, p. 101.
67. H. E. Katz, in J. L. Atwood, J. E. D. Davies and D. D. and MacNicol, eds., *Inclusion Compounds*, Vol. 4, Oxford University Press, Oxford, 1991, p. 391.
68. F. P. Schmidtchen, in F. Vögtle and E. Weber, eds., *Biomimetic and Bioorganic Chemistry II, Top. Curr. Chem.*, Vol. 130, Springer, Berlin-Heidelberg, 1986, p. 101.
69. E. Graf and J.-M. Lehn, *J. Am. Chem. Soc.* **98**, 6403 (1976).
70. F. Schmidtchen and G. Möller, *J. Chem. Soc., Chem. Commun.*, 1115 (1984).
71. P. D. Beer, N. C. Fletcher, A. Grieve, J. W. Wheeler, C. P. Moore, and T. Wear, *J. Chem. Soc., Perkin Trans 2*, 1545 (1996).
72. B. Dietrich, J. Guilhem, J.-M. Lehn, C. Pascard, and E. Sonveaux, *Helv. Chim. Acta* **67**, 91 (1984).
73. W. M. Hosseini and J.-M. Lehn, *Helv. Chim. Acta* **69**, 587 (1986); J.-M. Lehn, R. Müric, J.-P. Vigneron, I. Bkouche-Waksman, and C. Pascard, *J. Chem. Soc., Chem. Commun.*, 62 (1991).
74. A. M. Echavarren, A. Galón, J. de Mendoza, A. Salmeron, and J.-M. Lehn, *J. Am. Chem. Soc.* **111**, 4994 (1989).
75. G. A. Jeffrey and W. Saenger, *Hydrogen Bonding in Biological Structures*, Springer, Berlin-Heidelberg, 1991.
76. J. D. Watson and F. H. C. Crick, *Nature* **171**, 737 (1953).
77. A. D. Hamilton, in G. W. Gokel, ed., *Advances in Supramolecular Chemistry*, Vol. 1, JAI Press, Greenwich, 1990, p. 1.
78. D. A. Bell and E. V. Anslyn, in Ref. 12, Vol. 2, p. 439.
79. F. Garcia-Tellado, S. Goswami, S.-G. Chang, S. J. Geib, and A. D. J. Hamilton, *J. Am. Chem. Soc.* **112**, 7393 (1990).
80. J. C. Adrian and Jr., C. S. Wilcox, *J. Am. Chem. Soc.* **111**, 8055 (1989).
81. J. Rebek, Jr., *Science (Washington DC)* **235**, 1437 (1987).
82. J. Rebek, Jr., *J. Mol. Recogn.* **1**, 1 (1988).
83. J. Rebek, Jr., in E. Weber, ed., *Molecular Inclusion and Molecular Recognition—Clathrates II, Top. Curr. Chem.* Vol. 149, Springer, Berlin-Heidelberg, 1988, p. 189.
84. J. Rebek, Jr., D. Nemeth, P. Ballester, and F.-T. Lin, *J. Am. Chem. Soc.* **109**, 3474 (1987).
85. T. W. Bell, P. J. Cragg, M. G. B. Drew, A. Firestone, A. D.-I. Kwok, J. Liu, R. T. Ludwig, and A. T. Papoulis, *Pure Appl. Chem.* **65**, 361 (1993).
86. T. W. Bell and J. Liu, *J. Am. Chem. Soc.* **110**, 3673 (1988).
87. A. D. Hamilton and D. Van Engen, *J. Am. Chem. Soc.* **109**, 5035 (1987).
88. S. C. Zimmerman, in Ref. 10, p. 71.
89. M.-P. Teulade-Fichou, J.-P. Vigneron, and J.-M. Lehn, *Supramol. Chem.* **5**, 139 (1995).
90. B. Odell, M. V. Reddington, A. M. Z. Slawin, N. Spencer, J. F. Stoddart, and D. J. Williams, *Angew. Chem.* **100**, 1605 (1988); *Angew. Chem. Int. Ed. Engl.* **27**, 1547 (1988).
91. B. L. Allwood, N. Spencer, H.-Shahriari-Zavareh, J. F. Stoddart, and D. J. Williams, *J. Chem. Soc., Chem. Commun.*, 1064 (1987).
92. C. S. Wilcox, N. M. Glagovich, and T. H. Webb, in D. Truhlar and C. Cramer, eds., *Structure, Dynamics, and Reactivity in Aqueous Solutions*, ACS Symposium Series **568**, American Chemical Society, Washington DC, 1994, p. 282.
93. M. L. Bender and M. Komiyama, *Cyclodextrin Chemistry*, Springer, New York, 1978.
94. J. Szejtli, *Cyclodextrin Technology*, Kluwer, Dordrecht, 1988.
95. Ref. 12, Vol. 3.
96. C. D. Gutsche, *Calixarenes, Monographs in Supramolecular Chemistry*, Vol 1, The Royal Society of Chemistry, Cambridge, 1989.
97. J. Vicens and V. Böhmer, eds., *Calixarenes—A Versatile Class of Macrocyclic Compounds*, Kluwer, Dordrecht, 1991.
98. A. Pochini and R. Ungaro, in Ref. 12, Vol. 2, p. 103.
99. F. Diederich, *Cyclophanes, Monographs in Supramolecular Chemistry*, Vol. 2, The Royal Society of Chemistry, Cambridge, 1991.
100. Ref. 12, Vol. 2.
101. A. Collet, in Ref. 10, p. 103.
102. A. Collet, in Ref. 12, Vol. 2, p. 325.
103. J. C. Sherman, *Tetrahedron*, **51**, 3395 (1995).
104. D. J. Cram, *Nature* **356**, 29 (1992).
105. D. J. Cram and J. M. Cram, *Container Molecules and Their Guests, Monographs in Supramolecular Chemistry*, Vol. 4, The Royal Society of Chemistry, Cambridge, 1994.

106. P. Timmerman, W. Verboom and D. N. Reinhoudt, *Tetrahedron* **52**, 2663 (1996).
107. Y. Murakami, J. Kikuchi, and T. Ohno, in G. W. Gokel, ed., *Advances in Supramolecular Chemistry*, Vol. 1, JAI Press, Greenwich, 1990, p. 109.
108. Y. Murakami and O. Hayashida, in Ref. 12, Vol. 2, p. 419.
109. J. Franke and F. Vngtle, in F. Vngtle and E. Weber, eds., *Biomimetic and Bioorganic Chemistry II, Top. Curr. Chem.*, Vol. 132, Springer, Berlin-Heidelberg, 1986, p. 135.
110. F. Diederich, *Angew. Chem.* **100**, 372 (1988); *Angew. Chem. Int. Ed. Engl.* **27**, 362 (1988).
111. F. Diederich, *J. Chem. Educ.* **67**, 813 (1990).
112. H. J. Schneider, *Angew. Chem.* **103**, 1419 (1991); *Angew. Chem. Int. Ed. Engl.* **30**, 1417 (1991).
113. D. A. Dougherty, in Ref. 12, Vol. 2, p. 195.
114. F. Vngtle, C. Seel, and P.-M. Windscheif, in Ref. 12, Vol. 2, p. 211.
115. D. B. Smithrud, E. M. Sanford, I. Chao, S. B. Ferguson, D. R. Carcanague, J. D. Evanseck, K. N. Houk and F. Diederich, *Pure Appl. Chem.* **62**, 2227 (1990).
116. J. SD. Brodbelt and C.-C. Liou, *Pure Appl. Chem.* **65**, 409 (1993).
117. A. Galón, D. Andreu, A. M. Echavarren, P. Prados, and J. de Mendoza, *J. Am. Chem. Soc.* **114**, 1511 (1992).
118. C. Seel and J. de Mondoza, in Ref. 12, Vol. 1, p. 519.
119. W. Verboom and D. N. Reinhoudt, in Ref. 12, Vol. 2, p. 495.
120. M. T. Reetz, in Ref. 12, Vol. 2, p. 553.
121. J. Canceill, A. Collet, J. Gabard, F. Kotzyba-Hibert, and J.-M. Lehn, *Helv. Chim. Acta* **65**, 1894 (1982).
122. J. Jazwinski, J.-M. Lehn, D. Lilienbaum, R. Ziessel, J. Guilhem, and C. Pascard, *J. Chem. Soc., Chem. Commun.*, 1691 (1987).
123. M. T. Reetz, C. M. Niemeyer, and K. Hamms, *Angew. Chem.* **103**, 1515 (1991); *Angew. Chem. Int. Ed. Engl.* **30**, 1472 (1991).
124. M. T. Reetz, C. M. Niemeyer, and K. Hamms, *Angew. Chem.* **103**, 1517 (1991); *Angew. Chem. Int. Ed. Engl.* **30**, 1474 (1991).
125. F. C. J. M. van Veggel, W. Verboom, and D. N. Reinhoudt, *Chem. Rev.* **94**, 279 (1994).
126. D. M. Rudkevich, W. T. S. Huck, F. D. J. M. van Veggel, and D. N. Reinhoudt, in L. Fabbrizzi and P. Poggi, eds., *Transition Metals in Supramolecular Chemistry*, Kluwer, Dordrecht, 1994, p. 329.
127. S. Anderson, H. L. Anderson, and J. K. M. Sanders, *Acc. Chem. Res.* **26**, 469 (1993).
128. E. Weber, *J. Org. Chem.* **47**, 3478 (1982).
129. J.-M. Lehn, *Pure Appl. Chem.* **52**, 2441 (1980).
130. P. K. Coughlin, J. C. Dewan, S. J. Lippard, E. I. Watanabe, and J.-M. Lehn, *J. Am. Chem. Soc.* **101**, 265 (1979).
131. J. Monod, J.-P. Changeux, and F. Jacob, *J. Mol. Biol.* **6**, 306 (1963).
132. J. Rebek, Jr., *Acc. Chem. Res.* **17**, 258 (1984).
133. J. C. Rodriguez-Ubis, O. Juanes, and E. Brunet, *Tetrahedron Lett.* **35**, 1295 (1994).
134. H.-J. Schneider and D. Ruf, *Angew. Chem.* **102**, 1192 (1990); *Angew. Chem. Int. Ed. Engl.* **29**, 1159 (1990).
135. M. Inonye, T. Konishi, and K. Isagawa, *J. Am. Chem. Soc.* **115**, 8091 (1993).
136. E. L. Eliel, S. H. Wilen, and L. N. Mander, *STEREOCHEMISTRY OF ORGANIC COMPOUNDS*, Wiley, New York, 1994.
137. D. J. Cram and J. M. Cram, *Acc. Chem. Res.* **11**, 8 (1978).
138. J. F. Stoddart, *Chem. Soc. Rev.* **8**, 85 (1979); see also, J. F. Stoddart, in E. L. Eliel and S. H. Wilen, eds., *Topics in Stereochemistry*, Vol. 17, Wiley, New York, 1987, p. 207.
139. M. Pietraszkiewicz and N. Spencer, *J. Coord. Chem.* **27**, 115 (1992).
140. D. J. Cram, R. C. Helgeson, L. R. Sousa, J. M. Timko, M. Newkome, P. Moreau, F. De Jong, G. W. Gokel, D. H. Hoffman, L. A. Domeier, S. C. Peacock, K. Madan, and L. Kaplan, *Pure Appl. Chem.* **43**, 327 (1975).
141. V. Prelog, *Pure Appl. Chem.* **50**, 893 (1978).
142. J. F. Stoddart, in R. M. Izatt and J. J. Christensen, eds., *Progress in Macrocyclic Chemistry*, Vol. 2, Wiley, New York, 1981, p. 173.
143. J. P. Behr, J. M. Girodeau, R. C. Hayward, J.-M. Lehn, and J.-P. Sauvage, *Helv. Chim. Acta* **63**, 2096 (1980).
144. J.-I. Hong, S. K. Namgong, A. Bernardi, and W. C. Still, *J. Am. Chem. Soc.* **113**, 5111 (1991).
145. Y. Murakami, O. Hayashida, T. Ito, and Y. Hisaeda, *Chem. Lett.*, 497 (1992).
146. H. Grosenick, V. Schurig, J. Costante, and A. Collet, *Tetrahedron: Asymmetry* **6**, 87 (1995).
147. T. D. Booth, D. Wahnnon, and I. Wainer, *Chirality* **9**, 96 (1997); see also, V. Davankov, *Chirality* **9**, 99 (1997).
148. T. H. Webb and C. S. Wilcox, *Chem. Soc. Rev.*, 383 (1993).
149. S. K. Chang and A. D. Hamilton, *J. Am. Chem. Soc.* **110**, 1318 (1988).
150. T. R. Kelly and M. P. Maguire, *J. Am. Chem. Soc.* **109**, 6549 (1987).
151. A. D. Hamilton and D. Little, *J. Chem. Soc., Chem. Commun.*, 297 (1990).
152. G. Deslongchamps, A. Galón, J. de Mendoza, and J. Rebek, Jr., *Angew. Chem.* **104**, 58 (1992); *Angew. Chem. Int. Ed. Engl.* **31**, 61 (1992).
153. J. Haseltine and T. J. Doole, in *Organic Synthesis: Theory and Applications*, Vol. 3, JAI Press, Greenwich, 1996, p. 85.
154. T. D. James, K. R. A. S. Sandanayake, and S. Shinkai, *Angew. Chem.* **108**, 2039 (1996); *Angew. Chem. Int. Ed. Engl.* **35**, 1910 (1996).
155. T. D. James, P. Linnane, and S. Shinkai, *J. Chem. Soc., Chem. Commun.*, 282 (1996).
156. Y. Shiomi, M. Saisho, K. Tsukagoshi, and S. Shinkai, *J. Chem. Soc., Perkin Trans. 1*, 2111 (1993).
157. T. D. James, K. R. A. S. Sandanayake, and S. Shinkai, *Nature* **374**, 345 (1995).
158. Y. Aoyama, in Ref. 12, Vol. 2, p. 279; see also, K. Kobayashi, Y. Asakawa, Y. Kikuchi, H. Toi, and Y. Aoyama, *J. Am. Chem. Soc.* **115**, 2648 (1993).
159. T. Schrader, *Angew. Chem.* **108**, 2816 (1996); *Angew. Chem. Int. Ed. Engl.* **35**, 2649 (1996).
160. H.-J. Schneider, *Angew. Chem.* **105**, 890 (1993); *Angew. Chem. Int. Ed. Engl.* **32**, 848 (1993).
161. G. Wulff, W. Vesper, R. Grobe-Einsler, and A. Sathan, *Makromol. Chem.* **178**, 2799 (1977); see also G. Wulff, in J. S. Siegel, ed., *Supramolecular Stereochemistry*, Kluwer, Dordrecht, 1995, p. 13.
162. G. Vlatakis, L. I. Andersson, R. Möller, and K. Moosbach, *Nature* **361**, 645 (1993).
163. J. H. G. Steinke, I. R. Dunkin, and D. C. Sherrington, *Advances in Polymer Science* **123**, 81 (1995).
164. G. Wulff, *Angew. Chem.* **107**, 1958 (1995); *Angew. Chem. Int. Ed. Engl.* **34**, 1812 (1995); see also, G. Wulff, in W. T. Ford, ed., *Polymeric Reagents and Catalysts, ACS Symposium Series*, Vol. 308, American Chemical Society, Washington, D.C., 1986, p. 186.
165. E. Weber, ed., *Molecular Inclusion and Molecular Recognition—Clathrates I and II, Top. Curr. Chem.*, Vols. 140 and 149, Springer, Berlin-Heidelberg, 1987 and 1988.
166. E. Weber, in Ref. 165, Vol. 140, p. 1.
167. J. L. Atwood, J. E. D. Davies and D. D. MacNicol, eds., *Inclusion Compounds, Vol. 1–3*, Academic Press, Inc., London, 1984; Vols. 4–5, Oxford University Press, Oxford, 1991.
168. F. Toda, in Ref. 12, Vol. 6, p. 465; see also, F. Toda, in Ref. 165, Vol. 140, p. 43 and D. Worsch and F. Vngtle, in Ref. 165, Vol. 140, p. 21.
169. E. Weber, C. Wimmer, A. L. Llamas-Saiz, and C. Foces-Foces, *J. Chem. Soc. Chem. Commun.*, 733 (1992); see also, E. Weber and C. Wimmer, *Chirality* **5**, 331 (1993).
170. Ref. 12, Vol. 6.
171. C. L. Bowes and G. A. Ozin, *Adv. Mater.* **8**, 13 (1996).
172. T. Bein, ed., *Supramolecular Architecture, ACS Symposium Series*, Vol. 499, American Chemical Society, Washington, DC, 1992.
173. A. Möller, H. Reuter and S. Dillinger, *Angew. Chem.* **107**, 2505 (1995); *Angew. Chem. Int. Ed. Engl.* **34**, 2311 (1995).

174. I. Dance, in G. R. Desiraju, ed., *The Crystals as a Supramolecular Entity, Perspectives in Supramolecular Chemistry*, Vol. 2, Wiley, Chichester, 1996.

175. J.-H. Fuhrhop and J. Kinning, *Membranes and Molecular Assemblies: The Synekinetic Approach, Monographs in Supramolecular Chemistry*, Vol. 5, The Royal Society of Chemistry, Cambridge, 1994.

176. L. Addadi, Z. Berkovitch-Yellin, I. Weissbuch, J. van Mill, L. J. Shimon, M. Lahav, and L. Leiserowitz, *Angew. Chem.* **97**, 476 (1985); *Angew. Chem. Int. Ed. Engl.* **24**, 466 (1985); see also I. Weissbuch, R. Popovitz-Biro, L. Leiserowitz, and M. Lahav, in Ref. 4, p. 173.

177. I. Weissbuch, L. Addadi, Z. Berkovitch-Yellin, E. Gati, S. Weinstein, M. Lahav, and L. Leiserowitz, *J. Am. Chem. Soc.* **105**, 6615 (1983).

178. Z. Berkovitch-Yellin, *J. Am. Chem. Soc.* **107**, 8239 (1985).

179. I. Weissbuch, L. Addadi, L. Leiserowitz, and M. Lahav, *J. Am. Chem. Soc.* **110**, 561 (1988).

180. E. Delamarche, B. Michel, H. A. Biebuyck, and C. Gerber, *Adv. Mater.* **8**, 719 (1996).

181. L. H. Dubois and R. G. Nuzzo, *Annu. Rev. Phys. Chem.* **43**, 437 (1992).

182. R. M. Crooks, O. Chailapakul, C. B. Ross, L. Sun and J. K. Schoer, in T. E. Mallouk and D. J. Harrison, eds., *Interfacial Design and Chemical Sensing, ACS Symposium Series*, Vol. 561, American Chemical Society, Washington, D.C., 1994, p. 104.

183. N. Kodakari, N. Katada, and M. Niwa, *Chem. Vap. Deposition*, **3**, 59 (1997).

184. C. M. Paleos and D. Tsiourvas, *Adv. Mater.* **9**, 695 (1997).

185. H. Koyano, K. Yoshihara, K. Ariga, T. Kunitake, Y. Oishi, O. Kawano, M. Kuramori, and K. Suchiro, *J. Chem. Soc., Chem. Commun.* 1769 (1996).

186. J.-M. Lehn, M. Mascal, A. DeCian, and J. Fischer, *J. Chem. Soc., Chem. Commun.*, 479 (1990).

187. C. T. Seto, J. P. Mathias, and G. M. Whitesides, *J. Am. Chem. Soc.* **115**, 1321 (1993).

188. R. Ahuji, P.-L. Caruso, D. Mubius, W. Paulus, H. Ringsdorf, and G. Wildburg, *Angew. Chem.* **32**, 1082 (1993); *Angew. Chem. Int. Ed. Engl.* **32**, 1033 (1993).

189. K. Kurihara, K. Ohto, Y. Honda, and T. Kunitake, *J. Am. Chem. Soc.* **113**, 5077 (1991).

190. Y. Ikeura, K. Kurihara, and T. Kunitake, *J. Am. Chem. Soc.* **113**, 7342 (1991).

191. C. M. Paleos, Z. Sideratou, and D. Tsiourvas, *J. Phys. Chem.* **100**, 13 898 (1996).

192. J. S. Nowick, J. S. Chen, and G. Noronha, *J. Am. Chem. Soc.* **115**, 7636 (1993).

193. J. D. Dunitz, in G. R. Desiraju, ed., *The Crystal as a Supramolecular Entity, Perspectives in Supramolecular Chemistry*, Vol. 2, p. 1; see also, C. Pascard, in G. Tsoucaris and co-eds., *Crystallography of Supramolecular Compounds*, Kluwer, Dordrecht, 1996, p. 127.

194. A. I. Kitaigorodskii, *Molecular Crystals and Molecules*, Academic Press, New York, 1973; see also, C. P. Brock and J. D. Dunitz, *Chem. Mater.* **6**, 1118 (1994).

195. M. Simard, D. Su and J. D. Wuest, *J. Am. Chem. Soc.* **113**, 4696 (1991); see also, J. Rebek, *Acc. Chem. Res.* **17**, 258 (1984) and E. Fan, C. Vicent, S. J. Geib, and A. D. Hamilton, *Chem. Mater.* **6**, 1113 (1994).

196. J. C. MacDonald and G. M. Whitesides, *Chem. Rev.* **94**, 2383 (1994).

197. J. Rebek, Jr., *Acta Chem. Scand.* **50**, 707 (1996).

198. M. Mascal, N. M. Hecht, R. Warmuth, M. H. Moore, and J. P. Turkenburg, *Angew. Chem.* **108**, 2348 (1996); *Angew. Chem. Int. Ed. Engl.* **35**, 2204 (1996).

199. A. Ulmann, *Ultrathin Organic Films*, Academic Press, Boston, 1991.

200. R. Kramer, J.-M. Lehn, and A. Marquis-Rigault, *Proc. Natl. Acad. Sci. USA* **90**, 5394 (1993).

201. D. L. Caulder and K. N. Raymond, *Angew. Chem.* **109**, 1508 (1997); *Angew. Chem. Int. Ed. Engl.* **36**, 1440 (1997).

202. J.-M. Lehn and A. Rigault, *Angew. Chem.* **100**, 1121 (1988); *Angew. Chem. Int. Ed. Engl.* **27**, 1095 (1988).

203. A. Collet, in Ref. 12, Vol. 10, p. 113.

204. J. Jacques, A. Collet, and S. H. Wilen, *Enantiomers, Racemates, and Resolutions*, Wiley, New York, 1981.

205. J.-M. Lehn, in Ref. 4, p. 307.

Edwin Weber
Technische Universität Bergakademie Freiberg Institut für Organische Chemie

NANOTECHNOLOGY

Molecular nanotechnology (also known as nanotechnology, molecular engineering, or molecular manufacturing) is the production of functional materials and structures in the 0.1 to 100 nm range (the nanoscale) by any of a variety of physical and chemical methods. These methods include nanolithography, direct atomic and molecular manipulation with nanoscale probes, biotechnological selection and production of useful nanomaterials, and chemical synthesis and self-assembly of functional molecules and molecular aggregates (1–3). Various materials produced by such means are being investigated for a wide variety of current technological applications; they are also being examined as functional elements and building blocks of larger structures. Nanoscale structures and devices may be constructed synthetically from their atomic or molecular constituents (the synthetic or "bottom-up" approach, also referred to as nanochemistry) (4–6), or by fabrication techniques that use methods to form small structures from larger ones (the reductive or "top-down" approach) (7,8). Some of the structures and devices that have resulted, and have been envisaged to be forthcoming, from nanotechnology are molecular electronic devices (eg, molecular wires, diodes, transistors, memories), sensors, disease-fighting agents (eg, microbicidal, anticancer), and assemblers (devices that manufacture a particular molecule or assembly of molecules) using nanoscale mechanical devices (eg, nanopumps, motors, gears, and bearings) (2,9–12). A natural analogy to nanotechnology can be found in the biosphere, wherein small and large molecules interact to form complex structures (including many of the structures and devices listed above) necessary for all the functions of living organisms (13,14).

Synthetic vs Reductive Technologies

Many of the devices that have thus far been envisioned as products of nanotechnology (eg, nanoscale environmental sensors, information processors, and actuators) cannot be produced by the large-scale microfabrication techniques currently in use; indeed, these techniques (eg, photolithography) have nearly reached the limits of miniaturization imposed by both practical and fundamental constraints. The commercial production of nanoscale electronic materials is limited by factors such as electron tunneling, quantum mechanical confinement effects, heat dissipation, and difficulty and cost of fabrication (7,15,16). Although there continue to be advances in the development of processing methods that embody the reductive approach, the further development of nanotechnology now hinges on the understanding and manipulation of physical laws and processes at the nanometer level, such as electronic, interatomic, and intermolecular interactions that can be manipulated to allow efficient assembly of nanostructures. These include covalent, electrostatic, van der Waal's, dipole, and π -electron interactions, as well as hydrogen bonding, electron delocalization, and solvation (eg, hydrophobic) and thermal effects (17,18).

Biological systems employ a variety of synthetic strategies and processes that make efficient use of these interactions to form highly ordered molecular (eg, proteins, DNA) and supermolecular units (eg, plasma membranes, ribosomes) that perform a wide variety of functions, including light harvesting, catalysis, environmental sensing, information storage and processing, locomotion, and self-replication (4,19–23). These synthetic processes include covalent bond formation and polymerization, molecular templating, and molecular self-assembly. The distribution and functional success of biological nanostructures in living organisms are an evolutionally directed balance between the internal stability of the nanostructures (by atomic and molecular interactions) and their responses to the external environment. Thus, biological systems form important existence theorems and design criteria for the production of functional, nanoscale materials through nanotechnology employing the synthetic approach to nanostructures.

In recent years, a wide range of synthetic strategies have been developed that have sought to control molecular and atomic interactions to produce functional materials through nanotechnology using the bottom-up principle. These strategies have ranged in approach from those in which biological processes and synthetic pathways are utilized directly or are mimicked, to those that have no analogy in the natural world. Examples of synthetic strategies that have utilized biological processes directly include the expression of proteins and nucleic acids that are designed for specific function [eg, molecular recognition (24), catalysis (25,26), therapeutics (27–30)] or to exhibit specific structural properties (eg, three-dimensional structure) (14,24). Biomimetic strategies include schemes for formation of functional bilayer membranes (31), thin organic films (32), and vesicles (33) from synthetic precursors, and template-directed crystallization and polymerization (34). Other nonbiological strategies include total synthesis of functional, nonnatural molecules and local probe-assisted construction of atomic and molecular structures (35–39).

An important nature-mimicking methodology involves the use of covalent synthesis followed by molecular self-assembly of the synthesized molecules (4,5). These molecules are generally small mono- or oligomers that interact with each other and with other kinds of mono- or oligomers to form thermodynamically stable, nanoscale structures. Several interactions can play an important role in such assembly, eg, electrostatic, van der Waal's, dipole, and π - π electron interactions, hydrogen bonding, and hydrophobic effects (18). Biological systems that use these noncovalent interactions to assemble large molecules (eg, proteins, DNA) into compact, functional nanostructures provide the basis for these endeavors.

Related Technologies

Molecular nanotechnology is thus a consequence of current and future research and development efforts involving many broad areas of study, including biotechnology (14,40), protein and nucleic acid structure (41–44), supramolecular chemistry (2,4,6,45,46), molecular recognition (24,40,47), self-assembly (4,5), interface and colloid science (48,49), zeolite and clay chemistry (50), electrochemistry (51,52), electronics micro- and nanofabrication (8,53–55), and local probe microscopy (37–39,56). This article focuses on the current state-of-the-art of methods that produce nanostructures and devices. These methods include nanolithography, atom manipulation by local probes, biotechnology, nanochemical synthesis, rational design, and supramolecular chemistry.

Existence Theorem: Nanobiology

Analogies: Structure and Function. Biology is replete with complex, functional nanoscale structures formed by directed synthesis and subsequent self-assembly of the component molecules (13,14,57). Initially, the small molecules are produced by covalent synthesis, with the larger, functional structures resulting due to many weak, noncovalent interactions that energetically overcome interactions with the solvent and the entropic advantages of disintegration of the ordered aggregates. The final structure, thus, represents a thermodynamic minimum, and incorrect subunits are rejected in the dynamic, equilibrium assembly. In this process, complementarity in shape and polarity provides the foundation for the association (binding) between components (eg, phospholipids, polypeptide chains, proteins, nucleic acids). Shape-dependent association based on the nonspecific van der Waal's and hydrophobic interactions are made stronger and more specific by hydrogen bonds and electrostatic interactions, as well as, in some cases, covalent bonds (eg, disulfides) (14). Often, positive cooperativity is displayed, ie, conformations of individual subunits change upon binding in such a manner that their affinity for other components of the final structure increases. Moreover, the amount of information required to execute the assembly of a particular structure (eg, a protein) is minimized by use of only a few types of molecules and a limited number of binding interactions (13).

For example, a polypeptide is synthesized as a linear polymer derived from the 20 natural amino acids by translation of a nucleotide sequence present in a messenger RNA (mRNA). The mature protein exists as a well-defined three-dimensional structure. The information necessary to specify the final (tertiary) structure of the protein is present in the molecule itself, in the form of the specific sequence of amino acids that form the protein (57). This information is used in the form of myriad noncovalent interactions (such as those in Table 1) that first form relatively simple local structural motifs (helix

and sheet structures associated through networks of hydrogen bonds); these motifs then tend to aggregate in ways that associate hydrophobic regions with one another and out of reach of water, and to place hydrophilic regions close to each other and to water. Thus, proteins self-assemble by two types of processes after synthesis: formation of relatively simple local structures from an unfolded polypeptide chain; and more complex structure-specific association (ie, associations involving some form of molecular recognition) of these local structures (57).

Table 1. Types of Bonds and Interactions that are Potentially Useful in the Engineering of Functional Nanoscale Materials

Bond type	Examples	
	Natural	Synthetic
covalent bonds	peptide (amide), disulfide (RSSR), and phosphodiester (DNA) bonds	palytoxin, oligomers of pep-tides, nucleotides, and thiophenes
electrostatic interactions	salt bridges in proteins, zinc fingers	metal-coordinated crowns
hydrogen bonds	nucleotide base pairs, amide hydrogen bonds in proteins	peptide cylinders, DNA ob-jects, replicating systems
hydrophobic interactions	hydrophobic cores of proteins; lipid–lipid and lipid–protein interactions	pockets within cyclodextrins and cyclophanes
van der Waals interactions	membranes, vesicles	Langmuir and self-assembled monolayers, lipid bilayers
aromatic π -stacking and charge transfer	nucleic acids	porphyrins

In addition to self-assembly of protein structures, in living systems the complex maneuvers needed to achieve properly folded tertiary structures are facilitated by the function of a pre-existing protein machinery, of which, the molecular chaperones are an illustrative example (58,59). Chaperones (eg, heatshock protein 70 (Hsp 70) and the chaperonin families of proteins) are proteins that bind to and stabilize an otherwise unstable conformer of another protein, and by controlled binding and release, facilitate its correct fate *in vivo*, be it folding to attain proper tertiary structure, oligomeric assembly, transport to a particular subcellular compartment, or disposal by degradation. Molecular chaperones do not contain steric information specifying correct folding of a particular protein; instead, they prevent incorrect interactions within and between polypeptides in nonnative forms, thus typically enhancing the yield but not the rate of folding reactions (59). Hence, molecular chaperones may be said to be the natural counterparts of assemblers and transporters envisaged as products of nanotechnology (10,12).

Functional complexes produced by the association of proteins are also common in biology, from relatively simple examples (the association of four hemoglobin molecules into a tetramer or six insulin molecules into a hexamer) to extremely complex ones (eg, the formation of ribosomes) (60). The ribosome is formed by the ordered self-assembly of 55 proteins and three strands of RNA (13). Another example is the pyruvate dehydrogenase complex, made out of a total of 60 protein molecules that generate a functional structure with a diameter of ~30 nm (13). It is probable that enzymes that catalyze sequential reactions are colocalized within cells (60). Such protein–protein associations (which may be referred to as quinary structure) can facilitate the catalysis of reaction pathways by producing intermediates in the vicinity of enzymes that act upon those intermediates. Thus, such quinary structures may be considered to be the biological analogue of molecular manufacturing units envisaged by nanotechnologists (10,12).

Channeling can be advantageous for cell function in several ways (60). First, it serves to protect unstable or scarce metabolites by maintaining them in the protein-bound state. Second, channeling may provide a metabolic advantage by maintaining concentration gradients. A third function may be to protect the solvation capacity of cell water. If all of the several hundred low molecular-weight solutes had to diffuse through the entire aqueous phase of a cell's interior in order to find enzymes and to support reasonable reaction fluxes, that water would lose its ability to dissolve all of those metabolites, as well as the pools of soluble macromolecules. Finally, the kinetic coupling that causes channeling could provide kinetic advantages in addition to maintenance of concentration gradients. If intermediates are constrained to be in the immediate vicinity of an enzyme complex, then their pool sizes can be quite low, even while their local concentrations are high. Thus, the smaller the pools that must fill as the pathway is getting underway, the shorter the time needed for that pathway to reach peak efficiency. The simplest demonstration of channeling is demonstrated by one pathway of two reactions catalyzed by a single, bifunctional enzyme, tryptophan synthase (60). A low molecular-weight intermediate (indole) is channeled between active sites connected by a 3-nm tunnel within the interior of this protein.

Analogies: Molecular Devices. Among the many examples in biology of complex nanoscale structures and devices, one of the most ubiquitous and versatile classes is the membrane proteins. Membrane proteins, ie, proteins that are associated with cellular and organellar membranes, serve myriad functions including recognition, adhesion, chemical triggering, ion and molecular transport, light harvesting, chemical and mechanical sensing, and actuation. The diversity of the structure and function of these proteins found in nature is staggering, and humans have attempted to mimic the function of some of these structures for a wide variety of uses, including chemical sensing (31,61), selective chemical transport (pumping) (62), and tissue engineering (63). The importance of membrane proteins to the overall function of an organism is underscored by the drastic consequences to the survival of the organism resulting from disorders in which a particular membrane protein is not functionally expressed because of damage or a genetic disorder.

An example that illustrates the exquisite nanoscale complexity of a membrane protein system is the glycoprotein (GP) Ib-IX-V complex found on the surface of human platelets: blood-borne cell-like protoplasmic disks whose primary function is arresting blood loss (hemostasis). The surface of a platelet is covered with several types of glycoprotein complexes, of which the GP Ib-IX-V is the second most abundant (after GP IIb-IIIa); typically, approximately 25,000 copies of the complex reside on the platelet plasma membrane. The complex performs several important processes that contribute to the hemostatic functions of platelets (*vide infra*), each of which is vital to the well-being of the human organism. Genetic defects resulting in the nonexpression of the complex result in the most common form of the bleeding disorder, Bernard-Soulier syndrome (64).

Figure 1 schematically depicts the intricate structure of the membrane bound GP Ib-IX-V complex. The complex is formed and stabilized by several of the processes and interactions described above (also see Table 1), including those associated with protein folding and assembly into membranes (hydrophobic interactions, hydrogen bonding, disulfide bridging) and those resulting in the quinary structure of the complex (ie, those responsible for the stabilization of the seven-protein structure). The complex is a major attachment site between the plasma membrane and the platelet membrane skeleton. Through its interaction with actin-binding protein, the complex is likely to modulate actin polymerization, and hence, activate the platelet in response to extracellular cues such as binding of plasma proteins (eg, von Willebrand factor, thrombin) and shear.

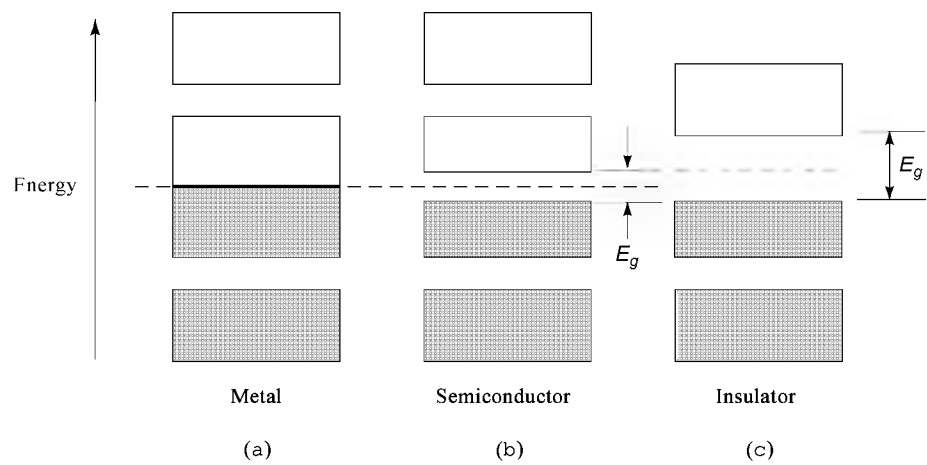


Fig. 1. The GP Ib-IX-V complex. The complex consists of seven transmembrane polypeptides denoted GP Ib α (mol wt 145,000), GP Ib β (mol wt 24,000), GPIX (mol wt 17,000) and GP V (mol wt 82,000), in a stoichiometry of 2:2:2:1. The hatched region represents the plasma membrane. The area above the hatched region represents the extracellular space; that below represents the cytoplasm. The complex is a major attachment site between the plasma membrane and the cytoskeleton. Two molecules associated with the cytoplasmic domain are depicted: a 14-3-3 dimer, which may mediate intracellular signaling, and actin-binding protein, which connects the complex to the cortical cytoskeleton and fixes its position and influences its function. The dark circles represent O-linked carbohydrate and the dark squares represent N-linked carbohydrate. Both types of carbohydrate are added after synthesis of the polypeptide chains (64).

Biochemical studies of the GP Ib-IX-V complex have identified several distinct functions of particular components of the complex that are related to the platelet's role in hemostasis. These include the binding of von Willebrand factor (vWf) that is immobilized on the vascular subendothelium (ie, the blood vessel wall), and the binding of thrombin. One of the most remarkable features of these interactions is that they are regulated so that they normally occur only when hemostasis is needed, ie, during bleeding. For example, it has been shown that GP Ib will bind vWf when vWf is immobilized on the vessel wall, and not when vWf is in solution in the plasma where it circulates along with the platelets. The ability of GP Ib to discriminate bound versus unbound vWf is thought to be related to the increased shear rates to which the platelet is subjected in flow near stationary surfaces. Thus, in addition to functioning as a molecular sensor and transducer (ie, transducing activation signals after vWf and thrombin binding), the GP Ib-IX-V complex is likely to serve as a mechanical sensor as well, whose molecular recognition capacity is activated only in the proper shear environment.

Another interesting class of membrane proteins that embody an example whereby a nanoscale device is produced by an organism and distributed into its environment for function comprises the pore-forming cytotoxins which are produced by a wide range of bacteria for the purpose of perforating the cells of other bacteria, or of host macro-organisms they are infecting. The gramicidin antibiotics are oligopeptides synthesized and secreted by certain bacteria for the purpose of permeabilizing, and hence killing, other bacteria. The ability of individual gramicidin peptides to assemble within lipid bilayers to form transmembrane pores has been utilized by Cornell and co-workers in synthetic molecular assemblies for biosensing (31).

The sensor (Fig. 2) comprises a supported, functionalized, lipid bilayer separated from a gold electrode by lipids substituted with spacers capable of forming gold-sulfur bonds. The membrane contains two types of gramicidin-based ion channels, one tethered to the gold, the other tethered to a hapten competitor and capable of diffusing laterally within the membrane. The membrane serves three purposes: (1) to anchor the recognition elements (antibody or Fab fragment and hapten competitor); (2) to allow diffusion of the hapten-tethered gramicidin; and (3) to prevent transport of ionic species between the analyte solution and the gold electrode. It may also serve to inhibit nonspecific interactions with other biological molecules in the solution to be analyzed. Upon binding the analyte, the released hapten-bound gramicidin can dimerize with the gold-tethered gramicidin to form a transmembrane pore that allows ionic transport to the gold surface. If a potential is applied, analyte binding is thus detected as a result of ion transport across the membrane. The fact that one analyte binding event can result in the transport of many ionic species across the membrane results in a large amplification effect during transduction (31).

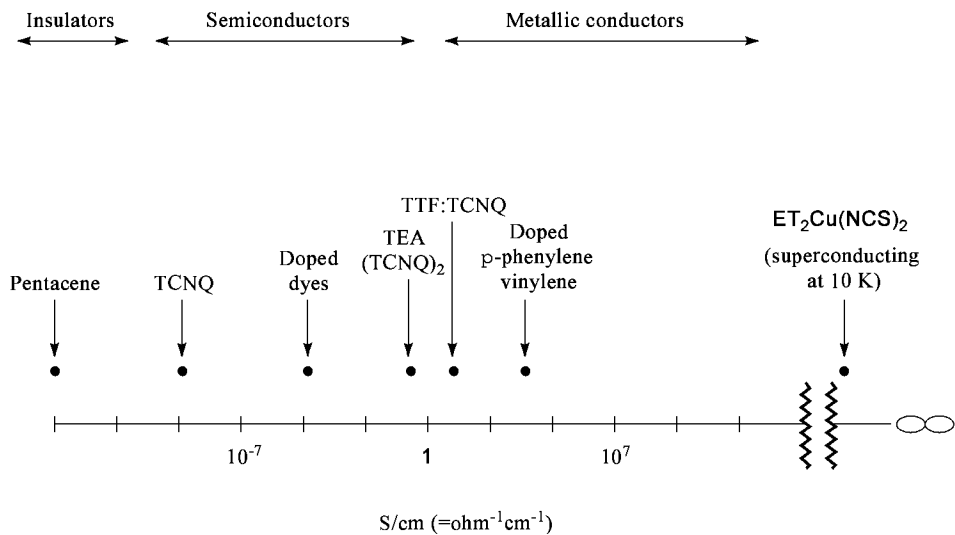


Fig. 2. (a) Schematic depiction of an example of a gramicidin ion channel-based biosensor that relies on competitive binding of an analyte and an analyte-analogue (bound hapten) on the surface of a supported lipid bilayer membrane. Analyte binding results in the lateral diffusion of the freed hapten-bound gramicidin (g1), dimerization of gold-bound gramicidin (g2) and hapten-bound gramicidin, and an increased ionic transport to the gold electrode. Adapted from Cornell and co-workers (31). (b) Schematic of the steps involved in the assembly of α -hemolysin, a heptameric transmembrane pore. Bacterial secretion of the single-chain monomers to their surrounding media *in vivo* results in binding of the monomers on cell membranes and subsequent pore formation. Four of the seven subunits are depicted in the final inserted pore. The intricate mushroom-shaped structure of the final pore structure has recently been elucidated by x-ray crystallography (68). Adapted from Bhaki and co-workers (66).

Staphylococcal α -hemolysin is another widely studied pore-forming toxin. It is used by infectious bacteria to perforate host animal cells by a mechanism that is distinct from that of gramicidin. Several aspects of the structure and function of this heptameric protein complex have been studied,

including its structure, assembly, mechanism of action, and transport properties (65–68). This and other similar pore-forming toxins are of great interest to biologists and biochemists not only because of their potent cytotoxic properties (66), but also because the proteins are a unique model system for studying the mechanisms by which transmembrane pores are formed (67). Of particular interest to the development of nanotechnological systems, this toxin has also been isolated and incorporated into synthetic lipid bilayers to develop pore-based chemical sensors (61). Bayley and co-workers have genetically modified the monomeric units of α -hemolysin so that their assembly into the heptameric pore complex occurs only in the presence of specific chemical [eg, metal ions (65), enzymatic (69)], or physical (eg, light) cues (70). Figure 2 gives a schematic representation of the steps involved in the assembly of the α -hemolysin transmembrane pore. The use of such assemblies for chemical sensing has the potential advantage that a large amplification effect can be obtained because the formation of one pore (eg, initiated by the binding of a target metal ion or other analyte) can result in the transmembrane transport of many indicator moieties (eg, ions) that can be measured (eg, potentiometrically) as an indication of the presence of analyte.

The two chemical sensing approaches discussed above (gramicidin-based and α -hemolysin-based) are examples of incorporation of functional biological molecules in nanotechnological devices, and also can be thought of as forming a putative basis for the biomimetic design of other complex nanostructures by synthetic means. An example of such biomimetic design is the synthesis by Ghadiri and co-workers of cyclic peptides that self-assemble in lipid bilayers (71). Figure 3 illustrates the structure of an example of these synthetic cyclic peptides and the ordered structure that they are thought to form upon self-assembly through hydrogen bond formation within a lipid bilayer. These structures are also currently being investigated for chemical sensing purposes both in lipid bilayer formulations and in other organic thin film architectures such as those based on self-assembled monolayers. The hydrogen-bonded structures are based on cyclic peptide subunits with alternating D- and L-amino acids that adopt flat ring-shaped conformations, and that allow for synthetic tuning of pore size and polarity, two characteristics that are likely to be important in the design of chemically-specific sensors. Similar synthetic cyclic peptides have been produced that self-assemble from solution into crystalline hexagonal arrays of nanotubules (72,73). By proper design of the cyclic peptides (eg, judicious choice of the type of peptides and of their sequence in the cyclic peptides), the nanotube arrays can be designed to display a variety of structural motifs (74).

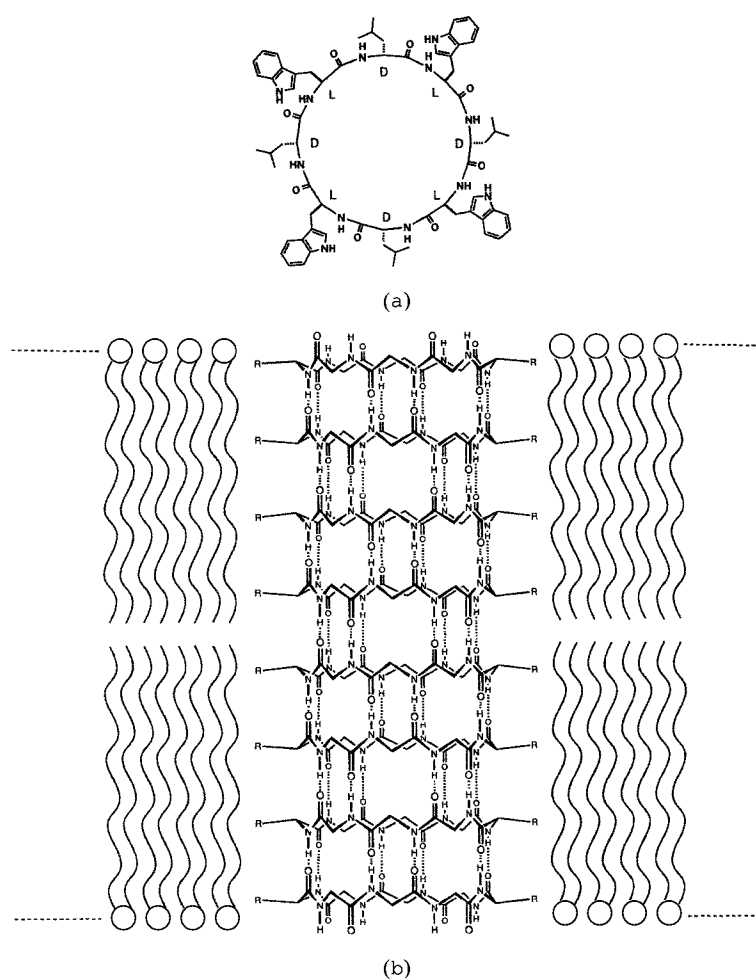


Fig. 3. (a) Chemical structure of a synthetic cyclic peptide composed of an alternating sequence of D- and L-amino acids. The side chains of the amino acids have been chosen such that the peripheral functional groups of the flat rings are hydrophobic and allow insertion into lipid bilayers. (b) Proposed structure of a self-assembled transmembrane pore comprised of hydrogen bonded cyclic peptides. The channel is stabilized by hydrogen bonds between the peptide backbones of the individual molecules. These synthetic pores have been demonstrated to form ion channels in lipid bilayers (71).

Courtesy of Dr. M. Reza Ghadiri.

Another interesting class of proteins which may prove to be an illustrative archetype for future nanotechnological devices, if not integral components of such, are a class of molecular devices collectively known as motor proteins. Motor proteins are enzymatic protein complexes whose catalytic function results in a distinct mechanical function; a variety of examples are known, including those that perform functions analogous to levers, rotary motors, pumps and springs (19). The structures and mechanisms of action of a variety of motor proteins are currently being investigated, including those of kinesin (75), myosin (76), actin (77), dynein (78), DNA helicases (79), and the bacterial flagellar motor (80).

Reductive Approaches

Conventional methods of microfabrication of integrated circuits and devices constitute the reductive (top-down) approach to the construction of micron- and submicron-scale structures (16). At the present time, the smallest features in commercial integrated circuits measure 0.35 μm across, and technologies for further reduction in size have succeeded in forming feature sizes as small as 0.18 μm (15,81). Further miniaturization, however, will require major technological breakthroughs in the processes underlying microfabrication, especially photolithography, the heart of microfabrication. The breakthroughs must not only allow further reductions in the size of the smallest features, but also be economically feasible to implement as manufacturing practices. Recent history of the semiconductor industry shows that, as the limits of a particular manufacturing technology are approached and surpassed, the economic consequences have been a spectacular increase in the cost of building new factories. At present, building a new silicon chip fabrication facility needs a capital layout of about \$1.5 billion. Although the size of the smallest possible features on the chip have shrunk by 14% every year, the price of

lithography equipment has risen at a rate of 28% annually (15). However, the technological barriers faced by the lithographers may not be completely insurmountable, nor may they be prohibitively expensive. In the following sections the limitations of current methods of microfabrication that prohibit their use for nanofabrication (fabrication of nanoscale structures and devices), new advances being made in current technologies that may be used in nanofabrication, and new reductive approaches to micro- and nanofabrication are discussed.

Limitations. The number of transistors present on a chip has doubled approximately every 18 months since the integrated circuit was first developed (a rate of increase predicted by Gordon Moore of Intel Corporation in 1960 that has become known as Moore's law). The main reason for this continuing decrease in the minimum feature sizes of transistors (and consequent increase in density of transistors on the chip) has been the development of photolithography, the prototypical reductive method (7,15,16,81–83). Photolithography creates patterns in layers of silicon, insulators, and metals to produce the integrated circuit. For the accurate reproduction of features onto the silicon wafer, the wavelength of light used must be at least as small as the smallest feature size (eg, feature resolution varies as the wavelength of light used and is inversely proportional to the aperture of the objective lens and the depth of focus (DOF), requiring innovative solutions in planarization technologies and mask design). Other issues related to further scaling down of integrated circuits include the effects of power supply and threshold voltage of the transistors, short channel lengths, thickness of the gate oxide, high electric fields, fluctuations in the number of dopant atoms, and interconnect delays (7,84). The most important limitation for further size reduction, however, remains the development of new photo- and other lithographic techniques.

Current Advances. New light sources are currently being developed, such as the krypton–fluoride ultraviolet laser (wavelength 0.248 μm for features as small as 0.25 μm) and excimer lasers (wavelength 0.193 μm for features below 0.2 μm) (81). But these technologies still need to overcome several obstacles before they can be implemented by the semiconductor industry. For example, for wavelengths below 0.2 μm, the current photoresists absorb so much light that throughput suffers. Also, the fused silica glass lenses used to demagnify the image absorb light and heat up, resulting in a degradation of the image. Problems associated with DOF become more acute, requiring further innovation in planarization technology. At the present time, these technologies are in the development phase, and it is expected that they will be ready for transfer to manufacturing facilities in the near future (15,81). New photoresist materials, based on the deposition and patterning of monomolecular layers (eg, self-assembled monolayers and Langmuir-Blodgett films), are also being developed to address the afore-mentioned issues of further reduction in the limits of photolithography; these materials and methods are discussed in "Molecular Self-Assembly."

Further reduction in feature size to achieve nanoscale structures by photolithography will necessitate the use of ever smaller wavelengths of light. For example, many manufacturers have explored the use of x-rays (wavelengths less than 1 nm) for the lithographic process for producing structures smaller than 0.1 μm (55,85). Use of x-rays, however, faces a host of obstacles to implementation (55,81,85). For example, commercial patterning tools for x-ray masks constitute a major challenge: they require masks made of materials such as gold or tungsten to absorb the x-rays. The features of gold masks must have very high aspect ratios (eg, 5) in order to absorb x-rays in opaque regions of the pattern that can have widths smaller than 100 nm. For optical and uv lithography, masks are typically four or five times larger than the actual image on a chip. They are focused onto the chip with demagnification to achieve a greater degree of definition in the lithography process. However, despite numerous efforts (86), no commercially feasible way to focus x-rays exists, resulting in the use of masks that are exactly the same size as the circuit. This means that at each of the 20 or more mask levels, an alignment tolerance of only a few tens of nanometers is required. This is a difficult goal to achieve. One approach to circumvent these problems involves using a high-powered laser to bombard a metal target to generate x-rays, which then illuminate a reflective mask. A series of mirrors then demagnify the image to its actual size on the silicon wafer (85).

A new technique developed by Brueck and co-workers for creating submicrometer feature sizes uses interference effects between two coherent laser beams along with multiple exposures (without the need for a mask) to form highly complex patterns in the photoresist (87,88). Linewidths as small as a quarter of the wavelength of the laser light used are achievable; for a readily available source such as an Ar⁺ laser (wavelength 364 nm, close to that of the mercury I-line), a minimum feature size as small as 90 nm is obtainable. Because the interference pattern does not vary with depth (ie, normal to the wafer) for distances of as much as several meters, the issue of DOF ceases to be a problem. It is also possible to project images as large as 30 × 30 cm². Indeed, interferometric lithography can be considered to be the limit reachable by the new mask-oriented methods being developed to extend photolithography for sub-0.2 μm features, such as phase-shift mask and off-axis illumination techniques (87,88).

Focused particle-based lithography, such as ion- (53), neutral atom- (89,90), and electron-beam lithography (91), are capable of achieving very high resolutions. For example, electron-beam lithography employs a focused beam of energetic electrons to draw lines directly on the photoresist (54,91). However, the method of writing each circuit feature separately is serial and inherently slow so that the technology cannot be used for simultaneous fabrication of many chips. This precludes the use of electron-beam lithography for direct writing of circuit patterns for ultralarge-scale integration. Instead, the technology is used to create the masks that are subsequently used in photolithography. To speed up the process of electron-beam lithography, methods are being explored to scan a broad electron beam across the entire chip by projecting the beam through an appropriate mask (in mimicry of photolithography) (91).

Even if new methods are developed for denser integrated circuits with smaller features, there are other areas of computing that present formidable challenges. For example, future computers would need data storage and handling capabilities in the hundreds of gigabits to terabit range (15,81). Existing compact disc- or magnetic drive-based storage devices cannot hold such vast amounts of information, nor can they deliver them to the ultra-high speed processors at a commensurate rate. One new approach to address these technical barriers is optical recording using holography (81). Holographic data storage uses lasers to write and read large blocks of data in a photosensitive material (eg, inorganic crystals such as lithium niobate, barium titanate, strontium barium niobate, and organic macromolecules such as bacteriorhodopsin) (81,92). By this method, storage densities of up to 1 terabit of data per cc of crystal are possible.

The most recent approach to reductive nanofabrication that can indeed construct nanoscale structures and devices uses microscopic tools (local probes) that can build the structures atom by atom, or molecule by molecule. Optical methods using laser cooling (optical molasses) are also being developed to manipulate nanoscale structures.

Atomic and Molecular Manipulation

Scanning Probe Microscopy. The scanning tunneling microscope (STM) can image and manipulate matter on the atomic scale (3,37–39,56,93). In general, when a small conducting probe (the tip, consisting of one or a few atoms, or a metal) is placed close (less than 10 nm) to the surface of a conducting substrate, an electronic current results under a suitable bias due to the overlapping of the electronic wave functions of the probe and the surface. Because this tunneling current is exponentially dependent on the separation of the probe from the surface, imaging resolutions of fractions of an angstrom can be obtained. These images reflect both the topography and the electronic structure of the surface. The STM can also be used to modify surfaces locally. As a result, individual atoms and molecules can be manipulated with atomic-scale precision. A variety of atomic manipulation processes have been demonstrated with the STM. They can be broadly classified as parallel processes, in which the atoms and molecules adsorbed on a surface are moved laterally with an STM tip, and perpendicular processes, wherein atoms and molecules are transferred from the tip to the surface or vice versa (3).

In parallel processes, eg, field-assisted diffusion and sliding, the bond between the surface and the adatom is never completely broken. Field-assisted diffusion of an adatom on the surface occurs due to the presence of the intense, inhomogeneous electric field between the probe tip and the surface, which gives rise to a potential gradient. Sliding of adatoms results because the tip of the STM always exerts a force on the adatom and this force can be tuned in terms of magnitude and direction by proper manipulation of the current, potential, or the position of the tip. This process is schematically shown in Figure 4, as employed by Crommie and co-workers (*vide infra*). The adsorbate is first located with the tip (STM in its imaging mode) and then the tip is placed near the adsorbate. The tip–adsorbate interaction is increased by increasing the tunneling current. The tip is then moved laterally across the surface at constant current to the desired destination, pulling the adsorbate along with it. The process is terminated by reducing the tunneling current (returning the STM to its imaging mode) leaving the adsorbate bound to the surface at the new location.

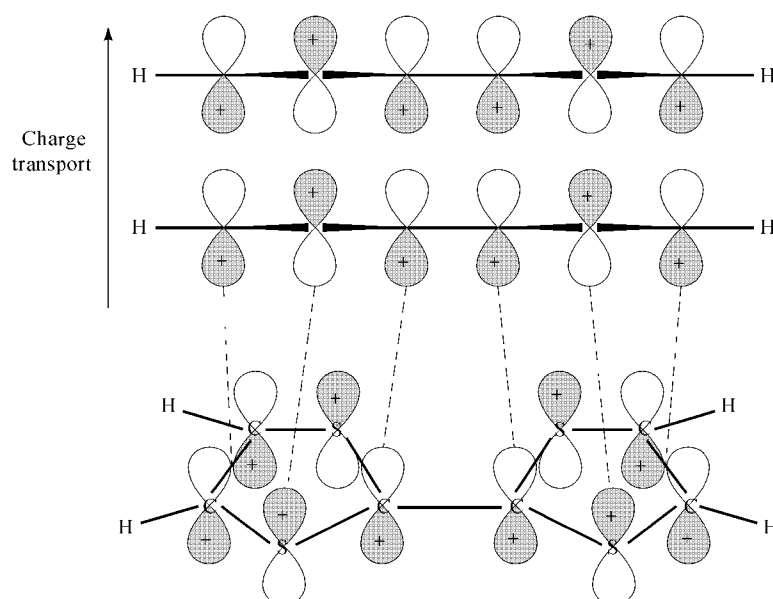


Fig. 4. Atom manipulation by the scanning tunneling microscope (STM). Once the STM tip has located the adsorbate atom, the tip is lowered such that the attractive interaction between the tip and the adsorbate is sufficient to keep the adsorbate "tethered" to the tip. The tip is then moved to the desired location on the surface and withdrawn, leaving the adsorbate atom bound to the surface at a new location. The figure schematically depicts the use of this process in the formation of a "quantum corral" of 48 Fe atoms arranged in a circle of about 14.3 nm diameter on a Cu(111) surface at 4 K.

Adapted from Crommie and co-workers (94).

The second class of atomic manipulations, the perpendicular processes, involves transfer of an adsorbate atom or molecule from the STM tip to the surface or vice versa. The tip is moved toward the surface until the adsorption potential wells on the tip and the surface coalesce, with the result that the adsorbate, which was previously bound either to the tip or the surface, may now be considered to be bound to both. For successful transfer, one of the adsorbate bonds (either with the tip or with the surface, depending on the desired direction of transfer) must be broken. The fate of the adsorbate depends on the nature of its interaction with the tip and the surface, and the materials of the tip and surface. Directional adatom transfer is possible with the application of suitable junction biases. Also, thermally-activated field evaporation of positive or negative ions over the Schottky barrier formed by lowering the potential energy outside a conductor (either the surface or the tip) by the application of an electric field is possible. Electromigration, the migration of minority elements (ie, impurities, defects) through the bulk solid under the influence of current flow, is another process by which an atom may be moved between the surface and the tip of an STM.

Several striking examples demonstrating the atomically precise control exercised by the STM have been reported. A "quantum corral" of Fe atoms has been fabricated by placing 48 atoms in a circle on a flat Cu(111) surface at 4K (Fig. 4) (94). Both STM (under ultrahigh vacuum) and atomic force microscopy (AFM, under ambient conditions) have been employed to fabricate nanoscale magnetic mounds of Fe, Co, Ni, and CoCr on metal and insulator substrates (95). The AFM has also been used to deposit organic material, such as octadecanethiol onto the surface of mica (96). New applications of this type of nanofabrication are being reported at an ever-faster rate (97–99).

Optical Manipulation. Laser beams provide another means of capturing individual atoms or molecules (89). When an atom is irradiated from both sides by laser light at a frequency slightly lower than the frequency at which the atom absorbs photons, then the atom loses some of its momentum. In particular, the laser beam propagating in a direction opposite that of the motion of the atom increases in frequency due to the Doppler effect, resulting in light absorption with subsequent isotropic emission (scattering). The light propagating in the same direction as the atom is not absorbed, so that the atom is pushed in a direction opposite its motion and slows down. By surrounding the atom with three sets of counterpropagating laser beams orthogonal to each other the atom can be cooled (ie, slowed) in all three dimensions (100). The first demonstration of this atom-trapping strategy resulted in the cooling of sodium atoms to 240 μ K. Because the light field acts as a viscous drag force, the combination of laser beams is known as optical molasses (89).

While near-resonant light exerts both scattering forces and dipole forces on single atoms, similar forces are also exerted on larger dielectric objects. A single laser beam can attract a dielectric object to its focal point through the dipole force, resulting in optical trapping; this has led to the development of a device called optical tweezers that can trap and manipulate small objects (ranging from large polymers to living cells) (89,101). Optical tweezers can handle live bacteria and other organisms without apparent damage (102). Even organelles within the cell can be manipulated without rupturing the cell wall (103). Optical tweezers are being used to study the mechanical properties of several biomolecular motors, including those associated with bacterial flagella and the contraction of muscles (104–106). On a smaller scale, single molecules of DNA have been manipulated and their elastic properties measured (107). New applications include the use of optical tweezers to measure the binding force of biomolecules to surfaces and to other biomolecules (108,109).

Optical trapping can also be used as a lithographic tool (90). For example, a combination of optical molasses and an optical standing wave have been used to focus a beam of neutral sodium atoms and deposit them in the desired pattern on a suitable substrate (eg, silicon). Pattern resolutions of the order of 40 nm with good contrast (up to 10:1 between the intended features and the surrounding unpatterned areas) and deposition rates of about 20 nm min were obtained (90).

Despite advances, it seems unlikely that the reductive approaches outlined above can, by themselves, reach the level of control, flexibility, discrimination, and versatility of atomic and molecular manipulation that will be needed to manufacture the molecular and supramolecular nanodevices envisaged to be the products of nanotechnology. Studies of biological nanodevices (eg, proteins) suggest that under proper conditions, atoms and molecules can assemble into functional nanoscale units that can carry out all the functions of life.

Synthetic Approaches

The synthetic (bottom-up) approach offers a level of control over the selection and placement of atoms and molecules that is ultimately much higher than that offered by other methods of large-scale microfabrication (eg, fabrication of integrated circuits) (2,4–6,17,22,46,49,110–112). Such synthesis can employ a variety of chemical methods utilizing some or all of the forces and interactions listed in Table 1 to produce nanoscale molecules. Most chemical synthetic reactions that produce large molecules (eg, collections of atoms that can act as nanodevices) generate polydisperse materials. These materials are mixtures of oligo- or polymeric chains of varying molecular weights. For nanotechnological applications, it is important to have synthetic strategies that yield compounds of uniform length, size, and shape; inhomogeneities can be detrimental to the designated function of the molecule. Thus, the specific objectives of synthesis are to discover and develop rapid and efficient methods for the precise control of composition, molecular weight, stereochemistry, aggregation, and placement of functional molecules. Four strategies are currently in use (either separately or in combination with one another) for the fabrication of large molecules (substances with molecular weights of a few hundred to a few million): biotechnological synthesis (14,113), sequential covalent synthesis (5,111), covalent polymerization (6), and molecular self-assembly (4).

In nature, complex functional molecules such as enzymes are produced by using the chemical synthetic approach outlined above (13). The molecules

are first formed by covalent synthesis. For example, proteins are formed under the direction of information contained in mRNA by the action of a variety of enzymes (protein catalysts) and cofactors. In particular, the polypeptide is formed by the directional joining of amino acids (brought to the site of the synthesis, eg, a ribosome, by a transfer RNA (tRNA) molecule) through amide bonds. The primary structure of the amino acids (ie, sequence) during synthesis is specified by the mRNA (a transcript of the original gene encoded in the DNA). Once the polypeptide is produced, it then undergoes many conformational changes (governed almost entirely by the sequence of the constituent amino acids) that reduces its size to a compact, native form that is the functional protein. These conformational changes are termed folding (41,57,59). As illustrated in Figure 5, protein folding proceeds by the formation, first, of local structural motifs (eg, α -helices, β -strands, and β -turns) by self-assembly. These motifs are formed as a result of formation of multiple hydrogen bonds between the side chains of amino acids that are far from each other in the linear sequence, but spatially close to each other. These structural motifs then self-assemble to form structural domains (eg, a bundle of α -helices, a barrel of β -sheets, and combinations of these) that are stabilized by hydrogen bonds and other noncovalent interactions (eg, the hydrophobic effect, electrostatic interactions, van der Waal's interactions). This secondary structure of the polypeptides aggregates further to form the tertiary structure (the final structure for the polypeptide chain) and quaternary structure (when more than one polypeptide chain aggregate to form the final protein structure). Often, covalent disulfide bonds and metal ions are employed to bridge the different structural motifs to provide additional structural stability. For example, a zinc-finger polypeptide coordinates a Zn^{2+} ion to provide the necessary stability to form a small protein, a task that would otherwise be impossible due to the small number of amino acids in the protein which precludes the formation of a large enough hydrophobic core (47). Additional chemical functionality, eg, carbohydrates, can also be incorporated in the protein structure through post-translational modification (13).

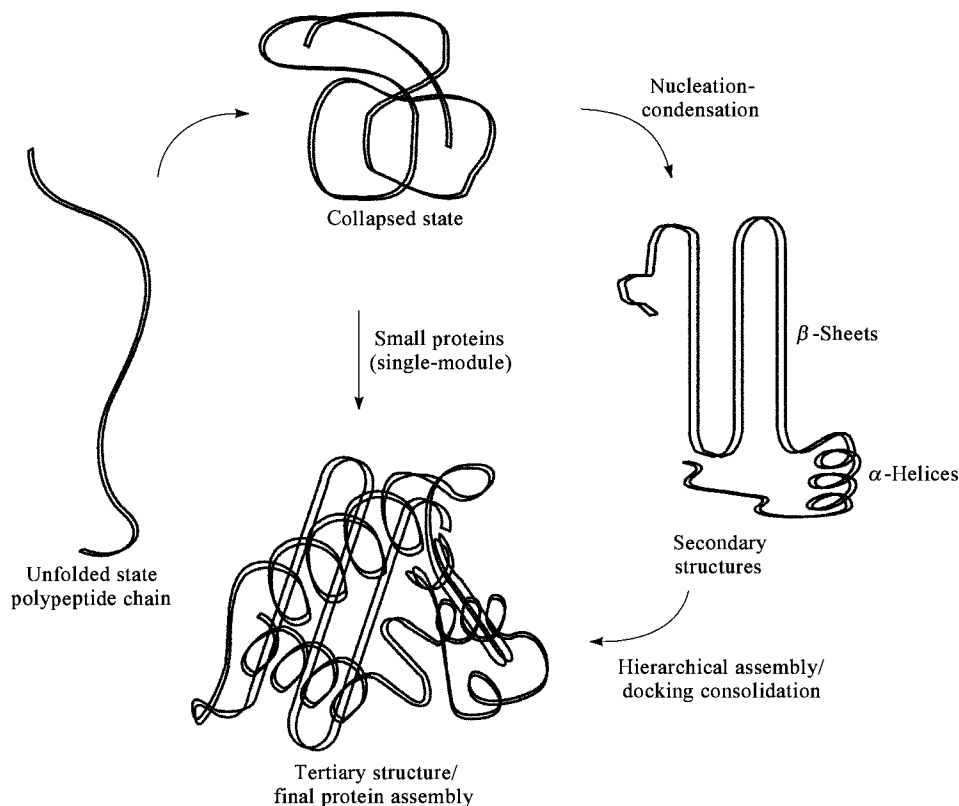


Fig. 5. Protein folding. The unfolded polypeptide chain collapses and assembles to form simple structural motifs such as β -sheets and α -helices by nucleation-condensation mechanisms involving the formation of hydrogen bonds and van der Waal's interactions. Small proteins (eg, chymotrypsin inhibitor 2) attain their final (tertiary) structure in this way. Larger proteins and multiple protein assemblies aggregate by recognition and docking of multiple domains (eg, β -barrels, α -helix bundles), often displaying positive cooperativity. Many noncovalent interactions, including hydrogen bonding, van der Waal's and electrostatic interactions, and the hydrophobic effect are exploited to create the final, compact protein assembly. Further structural consolidation may also occur by covalent bond formation (eg, disulfide bridges).

RNA is also capable of folding into specific shapes for ligand recognition and catalysis (44,114). Although there are only four naturally occurring bases available for the formation of tertiary structures through noncovalent interactions (while proteins have twenty different amino acids to choose from), myriad RNA shapes can be produced. The predominant noncovalent interaction involves the formation of hydrogen bonds. More importantly, metal ions have been shown to confer extraordinary stability to RNA and RNA fragments. For example, the overall tertiary structure of catalytic RNA (referred to as ribozyme) derived from the self-splicing intron of *Tetrahymena thermophila* results due to cooperative assembly of the RNA molecule upon uptake of at least three divalent metal ions (115). In particular, Mg^{2+} , Ca^{2+} , and Sr^{2+} can participate in general folding of the ribozyme tertiary structure, and Mg^{2+} is further involved in the catalytic activity of the ribozyme. The tRNA for methionine has a tertiary structure more stable than some of its secondary structural motifs in the presence of small concentrations of mono- and divalent cations (eg, Na^+ , Mg^{2+}) (44,114). This is primarily due to the creation of negatively charged pockets that have a high affinity for cations. On the other hand, the structural stability of a subunit of ribosomal RNA (rRNA) that binds ribosomal protein L11 and thiazoles is made possible by the presence of either Mg^{2+} or NH_4^+ cations. Such highly specific selectivity for the two ions implies that the rRNA coordinates or forms hydrogen bonds with these ions to form a tertiary structure that is rigid enough to distinguish between these ions and any others. It is important to note that the primary structure of the polynucleotides also determines the final, tertiary structure (similar to the case of polypeptides) (44,114).

Thus, protein and RNA folding studies form a fundamental paradigm for the design and synthesis of new, functional macromolecules with both final structure and function built into the primary structure of the macromolecule. Such synthetic strategies could be applied to the rational design of functional nanostructures, including drugs, sensing elements, photonic and electronic components, catalysts, and even mechanical devices. The current state of development of these methodologies is addressed in the following sections.

Biotechnological Synthesis. Molecular manufacturing of functional proteins within living systems follows the well-known path of transcription and translation of gene-coded information present in DNA (13). Typically, a transcript of the genes coding a particular protein is made. This transcript (mRNA) binds to complementary sections of two tRNA molecules. One tRNA contains an amino acid that is next in line to be attached to the growing polypeptide chain (which is attached to the other tRNA). The mRNA then translocates to bind another tRNA (bearing the next amino acid in the sequence) and so on. Biotechnological synthesis of new nanomaterials (eg, proteins) and biotechnological modification of living systems (eg, conferring specific therapeutic properties, new genetic traits) exploit the many ways in which this protein manufacturing machinery can be modified (13). In particular, the DNA of an organism can be altered in a specific manner, resulting in a modified mRNA, and consequently, a modified or new protein. The DNA can be altered by chemical synthetic methods followed by expression of the altered genes in a suitable organism (recombinant DNA technology), or by

selection and isolation of natural mutations.

Recombinant DNA technology provides a powerful tool for analysis, synthesis, and alteration of genes and proteins (13). It is based on the ability to rapidly synthesize polynucleotides with any sequence using nucleic acid enzymology (eg, using DNA polymerases, restriction enzymes or endonucleases, DNA ligases). The unique base-pairing attributes of the constituents of the DNA and the ability to express the modified or synthetic DNA in microorganisms and eukaryotic cells result in a powerful tool for the production of synthetic molecules with specific biological functions (13,116,117).

In general, a new DNA chain is made by joining fragments that have complementary cohesive ends produced by the action of a restriction enzyme (13). A DNA ligase then joins the ends of the DNA chains (provided they are within a double helix). Alternatively, tailor-made synthetic genes can be synthesized by sequential addition of activated monomers to a growing chain attached to a solid support, in a method similar to the Merrieffield synthesis (*vide infra*) (118). A further method involves the reverse transcription of mRNA to obtain the original DNA chain that encodes the protein that mRNA synthesizes. The DNA chains obtained in these ways are then inserted in vectors to obtain recombinant DNA. Examples of vectors include plasmids (naturally occurring circular DNA that act as accessory chromosomes), λ -phages, retroviruses, and adenoviruses (13). The recombinant DNA is introduced into producer cells (eg, *Escheria coli*) by direct uptake from the medium surrounding the cell, by injection into the host cell, or by use of a virus as a vector. The cells that contain the recombinant DNA are selected and cloned to obtain large quantities of cells with the ability to express the genes inserted into their genome. Gene expression in these cells results in the production of large quantities of useful proteins. This technology has now been commercialized to manufacture many naturally occurring, therapeutic protein drugs, including human insulin, erythropoietin, growth hormone, interferons, tissue plasminogen activator, and granulocyte colony stimulating factor (119).

It is also possible to carry foreign DNA into mammalian cells, for example, by use of retroviruses or by direct injection. For example, the gene for rat growth hormone was injected into fertilized mouse eggs, resulting in the production of giant mice that were almost twice as large by weight as their siblings with unchanged genome (120–122). Tumor-inducing plasmids (carried by the soil bacterium *Agrobacterium tumefaciens*) have been used to deliver foreign genes into plant cells, and represent promising candidates for exploring the genomes of plant cells and for modifying plants to improve their agricultural value (eg, disease resistance, increased yield) (13,123,124). Electroporation (application of intense electric fields that render the cell walls transiently permeable to large molecules), aerosol transport, and liposome delivery are other methods for transporting foreign DNA into cells (125–128).

The development of new, more efficient gene transfer protocols (eg, using retroviruses and adenoviruses) has allowed rapid progress in gene therapy (116,122,129–142). Here gene transfer involves the introduction of new genetic material into the cells of an intact organism in order to change its phenotype. Applications of gene therapy include substitution of missing proteins (due to genetic disorders, eg, deficiency of adenosine deaminase resulting in severe combined immune deficiency), increased synthesis of naturally occurring proteins (eg, cytokines in tumor cells for immune stimulation), expression of proteins not naturally occurring in humans (eg, production of anti-HIV antibody fragments in T cells), and targeted activation or deactivation of protein expression (eg, production of antisense RNA against HIV sequences). Expression of new genes in hematopoietic (primitive) stem cells is the focus of much research because it may provide a powerful tool for the cure of congenital genetic disorders (many cell lineages are derived from primitive stem cells throughout the life of an individual, so that genetic cures introduced into the stem cells can become nearly permanent).

The types of molecules synthesized by biotechnological techniques are restricted to those biomolecules whose structures can be encoded in the DNA of organisms capable of translating them into functional nanomaterials. Other types of molecules and nanomaterials can be synthesized by chemical synthetic approaches, such as covalent syntheses and molecular self-assembly of molecular units.

Covalent Synthesis. The first strategy for chemical synthesis employs elaborate and sophisticated methods for assembling atoms into molecules based on the general strategy of sequential formation of covalent bonds. The atoms can also be assembled into subunits that are then reacted to form more complex, designed molecules (convergent synthesis) (45). Covalent synthesis can be used to form molecules with well-defined composition, connectivity of atoms, and shape. This, however, requires enormous effort and energetic input when applied to very large molecules (or nanostructures), which hinders the use of this strategy alone to fabricate complex nanostructures that can interact with each other, with other nanostructures, and with their macroscopic environment.

Covalent synthesis of complex molecules involves the reactive assembly of many atoms into subunits with aid of reagents and established as well as innovative reaction pathways. These subunits are then subjected to various reactions that will assemble the target molecule. These reaction schemes involve the protection of certain sensitive parts of the molecule while other parts are being reacted. Very complex molecules can be synthesized in this manner. A prime example of the success of this approach is the total synthesis of palytoxin, a poisonous substance found in marine soft corals (35). Other complex molecules synthesized by sequential addition of atoms and blocks of atoms include vitamin B₁₂, the potentially anticancer KH-1 adenocarcinoma antigen, and epothilones A and B (36,143–145).

Covalent Polymerization. Other general methods of covalent synthesis include the controlled, stepwise addition of monomeric (or small oligomeric) units to a growing oligomeric chain that may be in solution or supported by covalent bonding to a solid substrate (5,45). The latter method, which includes oligopeptide synthesis, offers greater control over the efficient formation of precisely-defined polymers (ie, monodisperse polymers with controlled composition and length).

Families of linear conjugated oligomers such as α -oligothiophenes and α -oligo (thiophene vinylene)s, and cyclic conjugated oligomers of porphyrin and of thiophene (up to 200 membered rings) have been prepared (5). The long chains of oligo(α -thiophene ethynylene)s have been shown to act as molecular wires. Orthogonally fused conjugated oligomers have been synthesized that can act as molecular logic devices, although addressing the logic states is still a daunting problem. Precisely defined globular macromolecules (dendrimers, ie, structures that have cores with successively increasing branches of precise length and composition) have also been synthesized (146). Calixarenes, cyclodextrins, and cyclphanes are another class of cyclic oligomers that have been synthesized (5,147). These molecules can be used as hosts for conducting a variety of noncovalent chemistries, such as host-guest interactions within the hydrophobic cavity of the molecules and the formation of molecular scaffolds such as rotaxanes (147–150).

Synthesis of biopolymers such as oligo- and polypeptides using stepwise addition of amino acids to a growing chain of oligomers attached to a solid support (a method pioneered by Merrieffield (118)) has advanced to the stage of commercial manufacture. The synthesis involves the tethering of one end, eg, the carboxyl end, of an amino acid to an insoluble support (eg, polystyrene beads). The amino group of this tethered amino acid is then reacted with the carboxyl group of the next peptide in the sequence (other reactive functional groups in the amino acids are prevented from participating in unwanted reactions by use of protecting groups). The process is repeated to obtain a polypeptide tethered to the solid support. At the end of the synthesis, the carboxyl ester anchor is cleaved to separate the polypeptide from the solid support and the protecting groups are removed (13,118).

Combinatorial Chemistry. The development of stepwise methods of synthesis of polypeptides and polynucleotides has spawned a new methodology, combinatorial chemistry, for the production of new, functional molecules. Combinatorial chemistry is a way of rapidly generating a large number of closely-related compounds, followed by the screening of each of the compounds with respect to certain desired properties (eg, binding to a particular molecule) (151). A variety of methods of synthesis and screening utilize the combinatorial approach. These include parallel chemical synthesis and testing of multiple compounds (individually or in mixtures) in solution or on solid supports; and biotechnological or organism-based synthesis of biological oligomers coupled to methods of selection and amplification. In most applications, automated instrumentation has been developed that will produce a suite of compounds by covalent biochemical synthesis. These compounds may be small or large molecules (eg, oligopeptides and oligonucleotides of varying lengths) that are then screened by their ability to bind a biological molecule (eg, receptor or enzyme proteins). Once a compound that shows the desired binding properties is identified, a library of its variants can be produced, such that an optimized molecule can be obtained.

One of the many applications of the combinatorial approach is in the development of synthetic catalytic antibodies: molecules that mimic enzyme catalysts (25,26,152). A number of chemical transformations and reactions (some of them having rates of reaction comparable to those of enzyme-catalyzed counterparts) including kinetically disfavored and some nonenzymatic reactions, have been made possible by the use of catalytic antibodies. These catalysts are produced by screening against designed analogues of transition state species encountered in the reactions. Other applications of combinatorial chemistry include the production of catalytic RNA fragments (nucleic acids), and other small oligomers such as benzodiazepines and protease inhibitors (26,153,154). Molecules obtained from combinatorial libraries are finding extensive use as catalysts (25,152,155), medicinal drugs (28–30,113), and sensor elements (27,29,156,157). The combinatorial approach has also been applied to the discovery of new inorganic materials with specific electronic, magnetic, optical, chemical, and mechanical properties (158,159). For example, combinatorial arrays of films comprising different

combinations, stoichiometries, and deposition conditions of oxides of Ba, Bi, Ca, Cu, Pb, Sr, and Y were created and tested for superconductivity, resulting in the identification of mixed oxides of BiSrCaCu and YBaCu superconducting thin films (158) (see SEMICONDUCTORS).

Fullerenes. Many nanostructured, nonbiological polymeric systems find applications in catalysis and in separation of molecules. Buckminster fullerenes are closed-cage carbon molecules of the formula C_n , consisting of 12 pentagonal rings and any number (m) of hexagonal ones, such that $m = (C_n - 20) / 2$ (seven-membered rings are also possible) (160). The fullerene cages consist entirely of sp^2 -hybridized carbon atoms (resulting in electrophilicity) and act as giant closed-cage alkenes. Fullerenes exhibit fascinating optical, electrical, magnetic, and mechanical behavior (160–163). For example, reaction with good electron donors (eg, alkali metals) has lead to the preparation of crystalline fulleride phases that display superconductivity. Encapsulation of atoms (such as La, Ni, Na, K, Rb, Cs, Sc, U) within the fullerene cages has been demonstrated.

Tubular forms of fullerene (also called carbon nanotubes) are another example of a polymeric structure with many potential nanotechnological applications (164,165). For example, transition metals, carbides, and oxides have been encapsulated within carbon nanotubes forming nanowires (166,167). The transport of a single electron within a carbon nanotube has been demonstrated (168). Carbon nanotubes have high internal volume-to-surface area ratios and have been shown to store high volumes of hydrogen (169,170). Gadd and co-workers have demonstrated that carbon nanotubes can store argon gas at very high pressures (up to 60 MPa), thus acting as nanoscale gas cylinders (171). A single multiwalled carbon nanotube has been attached to the tip of a scanning probe microscope to yield an atomic manipulator of unprecedented dimensions. The nanotube tip is up to 1 μm in length and 5 nm in diameter (97,172). This tip can not only image features in trenches as small as 0.48 μm wide and 0.8 μm deep (which is not possible with the usual pyramidal tip), but also deposit a 40-nm dot of carbon at the bottom of the trench (172).

Template-Assisted Synthesis. Microporous (pore diameter <2 nm) phases with three-dimensional framework structures, eg, zeolites, are routinely synthesized by use of molecular units or hydrated alkali or alkaline earth cations as templates (50,173). Zeolites are crystalline inorganic oxides such as aluminosilicates that have uniform pores (size range 0.2 to 1.0 nm) with regular connectivity (173). Because of the constancy of pore characteristics of particular zeolites, they have been successfully used in commercial catalysis and separation systems, usually based on size and shape selectivity. The recent synthesis of silica-based mesoporous materials by the cooperative assembly of periodic inorganic and surfactant-based structures extends the range of molecular sieves (ie, separation systems based on the size and shape of molecules) into the regime of mesoporous materials (ie, 2 to 50 nm) (174). For example, researchers at the Mobil Oil Company used self-assembled molecular arrays as templates (instead of a single, solvated organic molecule or metal ion) and successfully synthesized the first mesoporous molecular sieves (MCM materials) with controlled pore sizes (between ~ 1.5 and ~ 10 nm) (175). These syntheses exploited the formation of micellar arrays of amphiphilic surfactant molecules whose noncovalent interactions with the inorganic polymerizable component (eg, alkoxysilanes such as tetraethoxysilane, TEOS) directed the formation of amorphous silica around them.

Organic ligands covalently bonded to the silica precursor (eg, an alkoxysilane) can be copolymerized with the inorganic component (eg, TEOS) to direct (ie, template) the formation of ceramic materials with very small (<1 nm) pores; unlike zeolites, these materials are amorphous (174,176). Typically, a nonporous hybrid material consisting of an organic template-containing precursor (eg, 3-methacryloxypropyl trimethoxysilane) and an inorganic polymerizable precursor (eg, TEOS) is formed by sol-gel processing (177), followed by the pyrolytic removal of the nonhydrolyzed organic template, leaving behind a material with pores that correspond to the size, connectivity, and perhaps, shape of the organic template ligands (176). The microporous material (pore diameter <1 nm) so obtained can not only exhibit molecular sieving properties similar to those of zeolites, but also allow greater ease and flexibility of processibility. For example, thin defect-free silica membranes can be produced with which sulfur hexachloride (kinetic diameter, $KD = 0.55$ nm) has been separated from helium ($KD = 0.265$ nm), nitrogen ($KD = 0.364$ nm), and propane ($KD = 0.45$ nm) (176).

The template-assisted synthetic strategies outlined above produce micro- or mesoporous structures in which amorphous or crystalline polymers can form around the organic template ligands (174). Another approach is the use of restricted spaces (eg, pores of membranes, cavities in zeolites, etc.) which direct the formation of functional nanomaterials within their cavities, resulting in the production of ultrasmall particles (or dots) and one-dimensional structures (or wires) (178). For example, in the case of polypyrrole and poly(3-methylthiophene), a solution of monomer is separated from a ferric salt polymerization agent by a Nucleopore membrane (linear cylindrical pores with diameter as small as 30 nm) (179–181). Nascent polymer chains adsorb on the pore walls, yielding a thin polymer film which thickens with time to eventually yield a completely filled pore. De-encapsulation by dissolving the membrane in CH_2Cl_2 yields wires wherein the polymer chains in the narrowest fibrils are preferentially oriented parallel to the cylinder axes of the fibrils. These polymeric heterocyclic nanotubes and nanowires have redox properties similar to those of parent, bulk polymers and their electronic conductivity is enhanced with respect to analogous bulk polymers (due to the preferential orientation of the polymer chains along the fibrils).

Nanoscale structures have also been fabricated by template-assisted lithography (eg, micro-contact printing and replica molding) using elastomeric (rather than rigid) organic polymer masters (182,183). The use of elastomers facilitates the easy separation of the master and the replica, thus minimizing damage to the small features on the replica as well as the master. The elastomeric master can also be shaped mechanically, allowing flexible control of the form and size of the features (184). For example, an elastomeric stamp made of poly(dimethylsiloxane) (PDMS) can be generated from an original master (with patterned features of metals, silica, silicon nitride, or photoresists) by cross-linking a viscous prepolymer after it has been poured onto the master. The stamp can be compressed mechanically to obtain patterned features smaller than those on the original master. The elastomeric stamp can then be re-replicated on a rigid organic polymer (eg, polyurethane). Polyurethane replicas with patterned features as small as 30 nm in width and 8 nm in height have been produced (185). Recently, PDMA stamps have also been made incorporating silicone fluid fillers that, when the filler is removed, yield patterned stamps with feature sizes smaller than those obtained without the filler (183). The methodology is potentially amenable to forming nanoscale, patterned organic layers with useful chemical, biochemical, magnetic, optical, and electronic functions.

A similar method of nanoscale patterning involves the use of an elastomeric resist film onto which the desired pattern is generated by compression molding at high temperature (ie, higher than the glass-transition temperature) (186,187). As a result of the compression, the film is thinned in the regions where the master pattern is present. Subsequent to controlled reactive ion etching, the thinned regions of the resist film are removed, leaving behind the thicker regions (which are partially thinned due to etching) and these form the desired pattern on the surface. This pattern can now be used as a resist layer. Poly(methyl methacrylate) (PMMA) is one candidate for this purpose, because it does not shrink or swell over large temperature and pressure ranges and is nonadhesive to silica. Metal (5 nm Ti and 15 nm Au) patterns with feature size of 25 nm have been fabricated by use of these resist templates in conjunction with metal deposition and lift-off (186,187).

Molecular imprinting is a technique by which molecular recognition capabilities are bestowed upon organic or inorganic polymeric systems through templating (174,188). The imprinting imparts structural information ("memory") of a particular molecule, by positioning the functional groups of the polymer (the local environment of the imprint molecule) in a specific geometric configuration that can then recognize the target (imprint) molecule. The polymeric materials are typically copolymers of methacrylates or vinylpyridines. Potential applications of imprinted polymers include catalysis, separations (eg, separation of chiral molecules), and chemical sensing (188).

Molecular Self-Assembly. As seen in "Reductive Approaches," reductive techniques, such as those currently in use in the microelectronics industry, can produce structural features smaller than about 200 nm, although at great economic cost. The use of proximal probes and other nanomanipulative techniques can be considered to be a hybrid of the reductive lithographic techniques and the synthetic strategies of assembling functional nanostructures atom by atom, or molecule by molecule. The organization of nanostructures and devices by the self-assembly of the component atoms and molecules, a ubiquitous phenomenon in biological systems, forms the noncovalent synthetic approach to nanotechnology.

In this approach, well-defined subunits (small molecules similar to, eg, nucleotides) are first formed through covalent synthesis. Second, these subunits aggregate with themselves or with other subunits through covalent or noncovalent (or both) interactions to form large, stable, structurally defined assemblies (eg, large molecules similar to RNA and proteins, and supramolecular assemblies such as DNA, ribosomes). To realize this synthetic approach, it is necessary to understand and control the noncovalent connections within and between molecules and to understand and overcome the possible unfavorable entropy increase involved in bringing many molecules together in a single, ordered aggregate (4,147). For the final supramolecular structure to be stable and to have a well-defined shape, the noncovalent connections must be collectively stable. The strengths of individual noncovalent bonds or interactions are small (eg, the strength of van der Waal's interaction is less than 4 kJ/mol, that of hydrogen bonds is of the order of 2 kJ/mol) relative to typical covalent bonds (150–400 kJ/mol) and comparable to thermal energies ($RT = 2.5$ kJ/mol at 300 K) (17). Therefore, molecules must be stabilized by many noncovalent interactions (ie, large complementary areas of molecular surface in interacting molecules must be in noncovalent contact).

The following sections contain a review of many of the varied synthetic systems that have been developed to date utilizing noncovalent interactions to form assemblies of molecules. These sections are loosely demarcated according to the most important type of noncovalent interactions utilized in conferring supramolecular order (ie, van der Waal's interactions, electrostatic interactions, and hydrogen bonds). For extensive reviews, see References 1,2,4–6,22,46,49,110–112. Finally, the development of self-assembling, self-replicating synthetic systems is noted.

van der Waal's Interactions. Langmuir mono- and multilayers of amphiphilic molecules (ie, molecules with a long hydrocarbon chain and a hydrophilic head-group) formed at the air–water interface are the prototypical example of a synthetic system based on van der Waal's interactions (189). These layers can form ordered, solid-like phases that can be transferred onto a variety of substrates (by physisorption), including glass, silicon, mica, and metal surfaces. Such films, called Langmuir-Blodgett (LB) films, retain their structural order and can display two-dimensional crystallinity. A variety of amphiphilic molecules have been investigated depending on the projected use of their LB films. In general, the molecules comprise a hydrophilic head group and a long hydrocarbon tail, that may be separated by a central group and or a linking group that provides the desired properties to the LB film. Common hydrophilic groups include quaternary ammonium, carboxylic acid, alcohol, methyl ester, and ethyl ester. Examples of central groups include biphenyl and azobenzene. LB films have also been fabricated from amphiphilic derivatives of porphyrins, phthalocyanines, polymerizable materials (eg, long chain acids containing vinyl derivatives and diacetylene groups), rod-like polymers (eg, synthetic polyglutamates), preformed polymers (eg, polyimides), and amphiphilic molecules incorporating electrically conductive materials (eg, the electron acceptor tetracyanoquinodimethane, TCNQ, and the electron donor tetrathiafulvalene, TTF) (189).

One of the foremost potential applications of LB films is in the generation of second harmonic radiation at optical frequencies. Noncentrosymmetric films are fabricated by using multilayers containing alternating monolayers. One monolayer contains chromophores with high second-order optical susceptibility and the other having no nonlinear optical properties (190). The LB technique has also been applied to fabricate noncentrosymmetric structures having pyroelectric properties (191). Multilayered LB films incorporating charge-transfer salts (CT salts such as TTF, TCNQ) as central groups have been prepared that display electrical conductivity comparable to that of not only pure crystals of the CT salts, but also metals (192). Even superconductivity has been observed in LB films containing the central group metal-bis-(4,5-dimercapto-1,3-dithiole-2-thione), where the metal is Ni or Au. LB film-based field-effect transistors have been fabricated that exhibit current–voltage characteristics typical of inorganic (eg, Si-based) transistors (192).

A second class of monolayers based on van der Waal's interactions within the monolayer and chemisorption (in contrast with physisorption in the case of LB films) on a solid substrate are self-assembled monolayers (SAMs). SAMs are well-ordered layers, one molecule thick, that form spontaneously by the reaction of molecules, typically substituted-alkyl chains, with the surface of solid materials (193–195). A wide variety of SAM-based supramolecular structures have been generated and used as functional components of materials systems in a wide range of technological applications ranging from nanolithography (196,197) to chemical sensing (198–201).

A prototypical class of SAMs that has been widely investigated is derived from the reaction of one or more substituted alkanethiols ($\text{HS}(\text{CH}_2)_n\text{R}$, where typically $n = 8-20$ and R represents one of a variety of diverse functional groups) with the surface of gold and other coinage metals (ie, silver, copper) (194,195). Figure 6a illustrates schematically the salient aspects of the structure of this type of SAMs, as inferred from the results of several investigations employing different analytical techniques such as x-ray photoelectron spectroscopy (202), polarized external infrared spectrometry (203), electron and helium diffraction (204), and x-ray reflectivity (205). These SAMs are easily prepared from solutions or vapors of the alkanethiols and are stable in common solvent environments, structurally well-ordered, and easy to modify synthetically (194,195). Gold supports for these SAMs can be easily fabricated by evaporation on a variety of substrates and can be thin enough to allow transmission optical microscopy, optical spectroscopy, and surface acoustic wave spectroscopy (201,206). These SAMs on gold also lend themselves to investigation by electrochemical methods and surface plasmon resonance (SPR) spectroscopy and microscopy and hence, have been exploited as platforms for affinity-based chemical sensors (198,199,201).

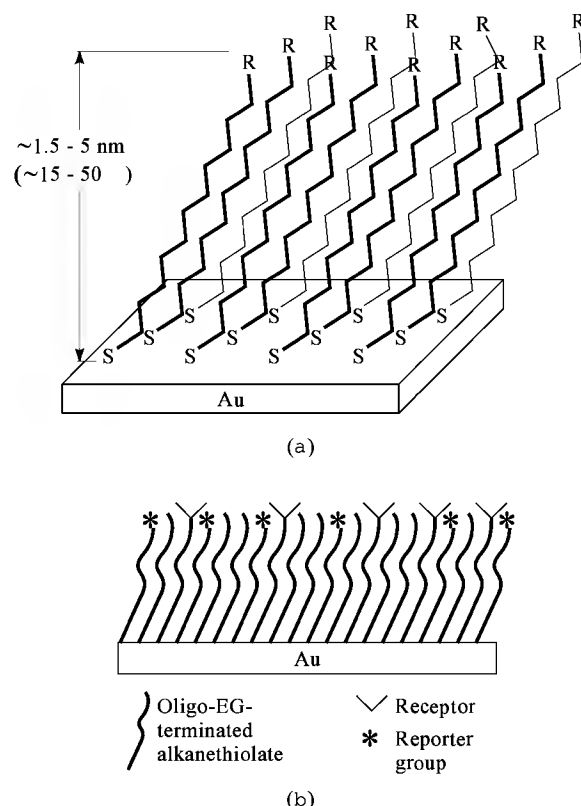


Fig. 6. (a) Schematic representation of the structure of a SAM derived by chemisorption of $\text{HS}(\text{CH}_2)_{11}\text{R}$ from solution onto a gold(111) surface (202,204,205,218). For simple terminal functional groups (eg, $\text{R} = \text{CH}_3$) the alkyl components of these SAMs are generally found to adopt a fully transextended configuration that is tilted from the surface normal by $\sim 30^\circ$. Note that this structure results in a surface whose chemistry is determined primarily by the identity of the terminal functional groups, R. (b) Schematic depiction of a modular SAM for use in biosensing applications. The illustration shows a three component mixed SAM of alkanethiolates terminated with a receptor moiety, a reporter moiety, and an oligo(ethylene glycol) moiety. A reagentless biosensor is formed by the specific interaction (binding) of an analyte (eg, a toxin, metabolite, or physiological indicator) with the receptor moiety (eg, a protein fragment, nucleotide, or synthetic receptor) which perturbs an externally-measurable chemical or physical property of the reporter group (eg, a fluorophore). The oligo(ethylene glycol) moieties serve to eliminate nonspecific interactions between the sensor surface and other components (eg, proteins or cells other than the analyte) of the solution to be analyzed.

A driving force for the development of this class of molecular assembly has been the ease with which these SAMs can be used to obtain different

interfacial chemistries by variation of the terminal functional group, R, and by forming multicomponent SAMs (eg, with two types of terminal functional groups R and R') (207). This property makes them ideal candidates as model systems for the investigation of interfacial chemical and physical phenomena. Further versatility in function can be obtained by forming patterned SAMs, in which particular areas of the surface contain one type of SAM (eg, with a particular terminal functional group, R) and other areas of the surface contain a different type of SAM (with R') or no SAM at all. Several methods have been developed for fabricating microscopic and nanoscopic patterns in SAMs of alkanethiolates on gold including: (1) use of polymeric "rubber stamps" to deliver reactive alkanethiols to many different areas of a gold surface simultaneously by micro-contact printing (208); (2) use of uv-lithography to oxidize thiolate groups to noncoordinating sulfonate-species that can be displaced by other alkanethiols (209); (3) use of micropens to deliver and to derivatize alkanethiols directly in specific areas of a gold surface (210); (4) micromachining of the gold to remove a pre-existing SAM followed by formation of a different type of SAM in the newly exposed gold regions (211); (5) use of local probes such as those used in scanning tunneling microscopy and atomic force microscopy to remove preexisting SAMs (212); and (6) electrochemical stripping of SAMs and blocking of SAM formation on gold microelectrodes (200). SAMs of alkanethiolates on gold or other substrates can also be patterned by electron- or ion-beam lithography (213). These methods for patterning SAMs have been used in a variety of experimental applications for micro- and nanofabrication and in the investigation of interfacial phenomena (see also LITHOGRAPHIC RESISTS).

From the standpoint of the use of SAMs as prototypical materials and functional elements in nanotechnological applications, both the ability to use these structures to control surface chemical properties precisely and the ability to form well-packed molecular layers, are of prime importance. Experimental systems based on SAMs have been devised to study phenomena in which the molecular constitution of the surface governs macroscopic to nanoscopic behavior of the surface. These phenomena include wettability (210,214), acid–base interactions (201,215), adhesion (216,217), corrosion (218), protein adsorption (209,219), biomolecular recognition, and cellular attachment and growth (208,220). Other nanotechnological applications have been in the area of material and nanostructures synthesis, and examples include the use of SAMs as resists for micro- and nanofabrication (196,197) and in templating the growth of nanoclusters, crystals and thin films (34,221). Of particular importance to nanostructure fabrication is the use of SAMs as ultrathin resist materials. Future reductive technologies requiring convenient resist materials for the production of materials with physical features of nanoscopic scale are likely to require nanoscopically thin resists for efficacy. In such cases, it may be that the use of self-assembling molecular films, whose thickness can be easily controlled by appropriate choice of the molecular constituents, and which have an inherent tendency to cover a surface conformally (and thus avoid defects), will allow the generation of resist materials for the high definition production of nanoscopically sculpted materials.

A number of efforts are underway to form SAMs for chemical sensing applications which mimic biological nanostructures (eg, membrane protein receptors) in that they exhibit modular function in which a recognition moiety (a receptor) is colocalized with a transducer moiety (a reporter) that signals the binding of an analyte to the receptor (see, for example, Fig. 6b). Another important function, especially for chemical sensing in biological milieu, that can be incorporated as another modular component in a mixed SAM, is the ability to resist nonspecific interactions of undesired biological components with the surface of the sensor. Figure 6b depicts schematically such a modular assembly for chemical sensing based on a multicomponent SAM of alkanethiolates on gold. Here the receptor is designed to bind specifically to a chemical analyte (eg, a toxin, a metabolite, a physiological indicator, or a drug), and the receptor group (eg, a fluorophore) is chosen so that its chemical properties (eg, fluorescence spectrum) is perturbed upon analyte binding to the receptor in such a manner that the perturbation can be reliably recorded by an external measurement technique (eg, fluorimetry). The third component of the SAM, here chosen to be an oligo(ethylene glycol) terminated alkylthiolate (eg, $-\text{S}(\text{CH}_2)_n(\text{OCH}_2\text{CH}_2)_m\text{OH}$), is incorporated to eliminate nonspecific adhesion of unwanted (ie, nonanalyte) species that may be present in the solution being analyzed (198,209,219,220). Poly(ethylene glycol) and other highly hydrated synthetic polymeric materials have been used as analogs of glycocalix structures on bacterial and other cellular surfaces that prevent nonspecific interactions of synthetic surfaces with biological components (209,219).

Another class of SAMs that have been widely used as functional nanostructures and in fabrication of nanostructures through lithography and templating are those formed by the reaction of alkyltrichloro- and alkyltrialkoxysilanes with the surface of silica and other oxide materials (222). Although the direct synthetic modification of substituted alkylsilanes is not as facile as that for corresponding alkanethiols, the modification of the surface chemistry of SAMs of alkylsiloxanes is possible, and these types of SAMs have been modified to incorporate a variety of simple functional groups and oligopeptides (223). SAMs of silanes can also be patterned by a variety of methods including through restriction of the area of derivatization with a patterned photoresist (210), or by direct uv or ion-beam lithography (221).

Recently, triblock copolymers having two aperiodic, coil-like blocks (styrene and isoprene, with a hydrophobic terminus) and one rod-like block (with a hydrophilic, phenolic terminus) have been synthesized (224). These triblock molecules self-assemble into mushroom-shaped nanostructures comprising about 100 molecules. The rodlike segments assemble in a parallel fashion with the two coil-like segments arranged in a space-filling mushroom shape. These structures can also form stacked layers. Such films, cast on glass and thermally cross-linked, adhere tenaciously to the substrate (the films could not be lifted by 6.7 M hydrofluoric acid, which can visibly etch glass). The films also exhibit nonlinear optical properties, eg, second-harmonic generation (224).

Electrostatic Interactions. Molecular host-guest complexes are composed of two or more molecules held together in unique structural relationships by forces other than fully formed chemical bonds (eg, covalent bonds). The host–guest relationship involves a complementary arrangement of binding sites in host and guest; the host has binding sites that converge in the complex and the guest has diverging binding sites within the complex. Electrostatic interactions (eg, metal–ligand coordination complexes) play an important role in the noncovalent stabilization of molecular and supramolecular structures (225,226).

Metal–ligand interactions can be strong, thus stabilizing the structure, while being kinetically labile, allowing reorganization (6,227). The concepts of self-assembly through metal coordination chemistry have been extended to construct a wide array of structures such as spherands, cavitands (molecular vessels), and carcerands (synthetic molecular cells), as well as planar structures such as ladders and bricks (226,228–232). Oligo-bipyridine ligands have been shown to spontaneously order themselves around an appropriate number of Cu(I) metal centers to give helical complexes (helicates), species that display many of the features associated with biological systems (22). For example, they show positive cooperativity; ie, the binding of the first Cu(I) ion facilitates the binding of the second, and so on. Bipyridine ligands with pendant nucleoside units around the Cu(I) centers have also been assembled (147). More recently, the self-assembly of triple helices (using Ni(II) ions) and the spontaneous formation of a cylindrical complex have also been demonstrated (147). Helical superstructures arise when the oligobipyridine ligands interact with a small, charged center that can impose a stereochemical preference on the system. These systems have been used in the construction of molecular racks, ladders, and grids (228). For example, a cyclic tetrameric macrocycle, ie, a molecular box, is formed when M(ethylenediamine) (NO_3)₂ (M = Pt, Pd) is treated with bipyridine.

The size-exclusion and ion-exchange properties of zeolites have been exploited to cause electroactive species to align at a zeolite–water interface (233–235). The zeolite thus acts as a template for the self-organization of electron transfer (ET) chains that may find function as biomimetic photosynthetic systems, current rectifiers, and photodiodes. An example is the three subunit ET chain comprising $\text{Fe}(\text{CN})_6^{4-}$ anion (which is charge-excluded from the anionic zeolite pore structure), $\text{Os}(\text{bipyridine})_3^{2+}$ (which is an interfacial cation due to size exclusion of the bipyridine ligand), and an intrazeolite cation (trimethylamino)methylferrocene⁺ (Fc^+). A cationic polymer bound to the $(\text{CN})_6^{4-}$ anion holds the self-assembled structure at an electrode. The triad acts as a current rectifier because the intrazeolite Fc^+/Fc^0 couple is unable to communicate directly with the electrode. Upon activation of the electrolytic cell comprising this triad with an alternating potential, electrons only pass from Fc^+/Fc^0 to the electrode as the $\text{Os}(\text{bpy})_3^{2+}/\text{Os}(\text{bpy})_3^{3+}$ couple is more oxidizing than Fc^+/Fc^0 . Current rectification in the opposite sense has also been realized with a $\text{Fe}(\text{CN})_6^{4-}$ – $\text{Ru}(\text{bpy})_3^{2+}$ –cobalticenium triad. By incorporating suitable light absorbing molecules into the chain, photodiodes have been constructed to resemble the donor–acceptor–acceptor triad in the photosynthetic reaction center of purple bacteria, in which extremely efficient and irreversible light-induced charge separation occurs.

Another synthetic strategy is based on self-assembly driven by molecular recognition between complementary π -donors and π -acceptors. Examples include the synthesis of catenanes and rotaxanes that can act as controllable molecular shuttles (6,236). The π -donors in the shuttles are located in the dumb-bell shaped component of the rotaxane and the π -acceptors in the macrocyclic component, or vice versa. The shuttles may be switched by chemical, electrochemical, or photochemical means.

Hydrogen-Bonded Assemblies. Specific, highly directional interactions such as hydrogen bonding have been used to achieve self-assembly (6,17,22). Hydrogen bonding depends on the structure of both the hydrogen-bond donor and acceptor, and their orientation relative to each other. For example, guanosine monophosphate spontaneously assembles *in vivo* to produce planar rings that can stack themselves in a cylindrical fashion (147). Each planar disk comprises four hydrogen-bonded guanine bases. This observation has led to numerous reports of similar molecules that can assemble by hydrogen bonding into planar and columnar aggregates (eg, folic acid, isoguanosine, dipyrindones) (147).

The amine functional group has been exploited to synthesize a variety of molecules that can form hydrogen bonds. For example, Whitesides and co-workers have demonstrated that cyanuric acid and melamine and their derivatives can assemble into myriad supramolecular structures (237,238). As shown in Figure 7, three melamine subunits can be attached to a central hub; the layer of melamines is repeated to obtain a three-tiered molecule **A**. The geometry of the molecule is such that it can bind nine molecules of the cyanuric acid derivative **B** to form a triply-stacked supramolecular aggregate (molecular weight ~6 kDa) stabilized by 54 hydrogen bonds. Similar to the amide functionality, carboxylic acids can also be employed to form hydrogen-bonded networks in aprotic solvents (147). For example, trimesic acid (1,3,5-benzenetricarboxylic acid) self-assembles into a sheet of hexagonal arrays. Two dimers of biphenyl-3,3'-dicarboxylic acid (which exists as two conformers syn and anti) can self-assemble to produce a cyclic aggregate that looks similar to the number eight. A porphyrin derivative containing two alkyluracil groups has been shown to coordinate a triaminopyrimidine derivative to give a supramolecular cage. Metallorganic compounds are another class of molecules that can form extended hydrogen-bonded complexes (147). For example, a 3 × 3 planar grid has been generated by the complexation by Ag(I) of 3,6-di-2-pyridylpyridazine (229).

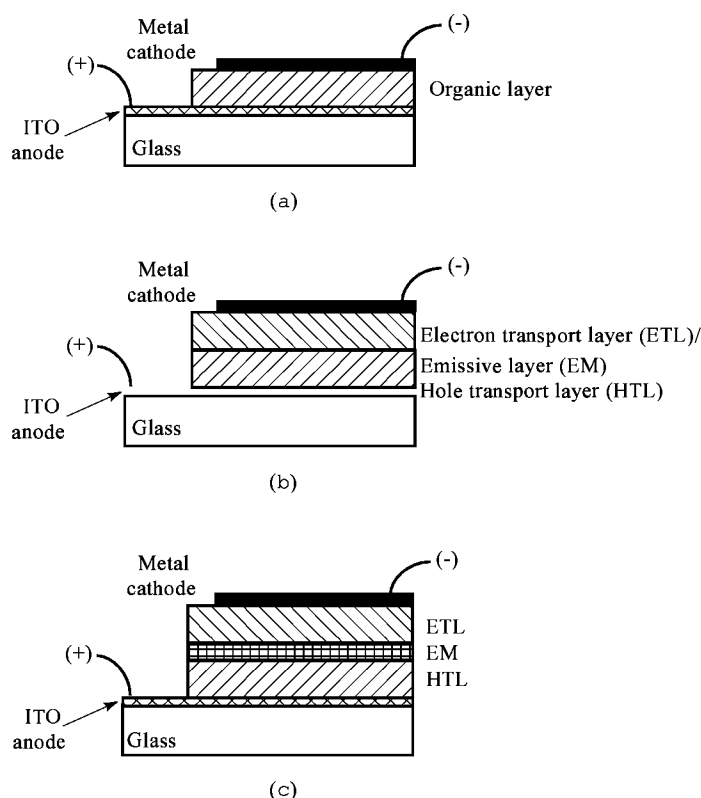


Fig. 7. Formation of a supramolecular aggregate composed of a compound containing nine melamine rings (the three-layered nonamelamine derivative **A**) and nine molecules of neoheptyliscyanurate (**B**). Of the 16 possible conformers that can result, two are shown: The first has the nine molecules of **B** arranged in three rosettes of three molecules each, stacked atop each other; in the second, the rosettes are staggered with respect to each other such that the rosettes in the first and third layers of **A** are aligned with each other, but not with the rosette in the second layer. The supramolecular assembly is stabilized by 54 hydrogen bonds.

Adapted from Mathias and co-workers (238).

Cyclodextrins and crown ethers also have the ability to form supramolecular aggregates with various shapes (eg, cyclic or helical) (148,239). For example, a β -cyclodextrin derivative containing four lipophilic arms was incorporated into a lecithin bilayer. The compound dimerizes in the bilayer to form an ion channel that enhances the rate of transport of Co^{2+} ions (by an order of magnitude) into the interior of the liposome, compared to an ion channel formed by 18-azacrown-6 (147). Synthetic cyclic peptides containing alternating D- and L-amino acids can self-assemble by hydrogen bonding into stacked rings with a central pore that can act as an ion channel (see Fig. 3). Ghadiri and co-workers have produced such an ion channel from a cyclic octapeptide with a pore size of 0.75 nm (71). The ion channel so-formed has transport properties similar to those of gramicidin A; indeed, the synthetic ion channel (nanotube) transports cations three times faster than gramicidin A. The peptide nanotube also displays selectivity for K^+ versus Na^+ . Recently, peptide nanotubes with 1.0 nm and 1.3 nm diameter pores capable of glucose transport were prepared (240).

Helical arrays are another structural motif that noncovalent interactions can generate (eg, DNA in living systems). Along with hydrogen bonds, the hydrophobic effect and metal ion coordination are important contributors to the formation and stability of such supramolecular complexes. For example, tartaric acid-based derivatives have been synthesized that can form either left- or right-handed helical superstructures through hydrogen bonding between complementary uracil–pyridine base pairs; the handedness of the helix is determined by that of the tartaric acid derivatives (147). Double helices have also been demonstrated. For example, two molecules of quinquepyridine (or sexipyridine) intertwine upon exposure to metal ions to produce a double helix. Ghadiri and co-workers have synthesized polypeptides that can self-assemble into a variety of helices upon coordination by metal ions (eg, Ru^{2+}) to produce three- and four bundle helices (241–244).

The self-assembly of synthetic DNA molecules also offers control of the structure of matter on the nanometer scale (245). The natural, linear, double-stranded DNA assembly invariably forms a right-handed double helix in compliance with the complementary rule that adenine always binds to thymine and that guanine always pairs with cytosine. Because the hydrogen bonding motif is well-known and easily varied, the possibility of designing synthetic sequences of nucleotides that self-assemble to form branched DNA molecules allows DNA to be engineered in three dimensions and serve, perhaps, as a molecular scaffold for placement of other functional molecules. This has been successfully demonstrated by the preparation of synthetic, branched DNA molecules that can assemble into two-dimensional lattices, three-dimensional cubes and truncated octahedra (245). Such DNA arrays and objects may be useful as macromolecular hosts, as scaffolds for molecular electronic components and for tethering catalysts, proteins, or drugs (245). The isomerization reactions of these DNA lattices (eg, from the normal right-handed structure to a left-handed structure) results in a torsional transition that, with proper choice of the parts of the array that are isomerized, can lead to controlled molecular motion, a key goal of nanotechnology.

Self-Replicating Systems. Recently, molecules have been synthesized that can catalyze covalent bond-making reactions by forming a noncovalently bonded superstructure, a maneuver that converts an intermolecular reaction into an intramolecular one. In general, in such systems, two

reaction channels exist: an uncatalyzed bimolecular reaction and a catalyzed ternary reaction. When the reaction product can serve as a template for its own formation and is capable of conserving and expressing its structure, a self-replicating process results (246–249). For example, the aminoadenosine derivative X is recognized by the pentafluorophenyl ester Y, and forms the complex XY (Fig. 8) (250). This complex is stabilized by the hydrogen bonds between the imide unit of Y and the adenine residue of X, and probably also by the π - π stacking interactions between the adenine base of X and the naphthalene moiety of Y. This arrangement allows the formation of a *cis*-amide bond between the amino function of X and the activated ester of Y. The *cis*-conformer of the product Z (*cis*-Z) rapidly isomerizes to the more stable *trans*-amide Z, opening up two new binding sites capable of binding X and Y. The binding of X and Y by Z is analogous to the formation of a ternary complex containing the reactants and the bisubstrate reaction template; in this case, moreover, the product of the reaction is Z. The ternary complex facilitates the reaction of X and Y to form the hydrogen-bonded complex ZZ. This dimer can dissociate to form two separate molecules of Z (typically, the noncovalently bonded dimer ZZ is in thermodynamic equilibrium with free Z). Thus, Z is self-replicating. In these and other systems, competition, reciprocal template effects, and mutation have also been observed (246–248).

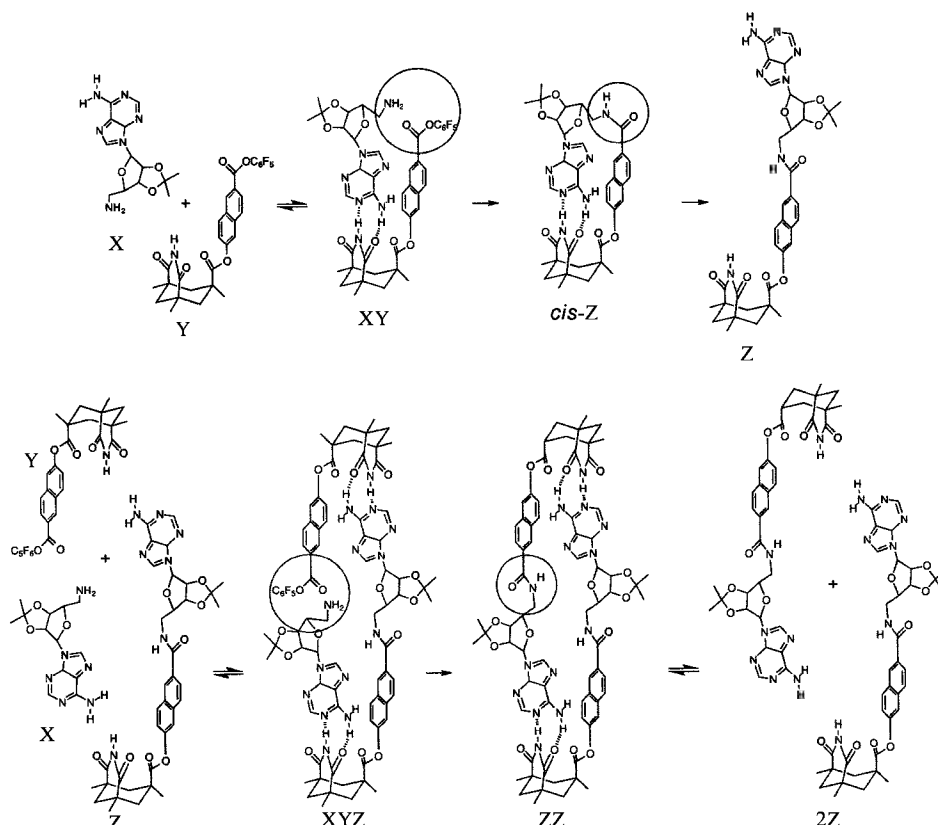


Fig. 8. Replication. The aminoadenosine X and the pentafluorophenyl ester Y form a hydrogen-bonded dimer XY, prior to reaction between the amine and the activated ester groups (shown in the circle). The reaction product is a *cis*-amide conformer *cis*-Z that isomerizes to the more stable *trans*-amide Z. The replicative process is catalyzed by the reaction product Z (also referred to as the template). First, a ternary complex XYZ is formed from X, Y, and Z.

The amide formation reaction (highlighted by the circle) leads to the production of a hydrogen-bonded dimer (ZZ) of the reaction product Z with the template Z. The dimer is in thermodynamic equilibrium with free template in the reaction medium.

Adapted from Nowick and co-workers (250).

Similar to the presumed earliest forms of single-celled life, nano- and microstructures have been created that point to a possible approach to the construction of a synthetic, living cell (251,252). This approach involves the use of self-assembled, self-reproducing vesicles of surfactant molecules that incorporate within their cavities the reactants and catalysts necessary for the self-replication of the molecules bearing the information (in their structure) for the replication of themselves as well as the encircling vesicles. The first steps of these syntheses have been demonstrated. At the present time, fragments of catalytic RNA have been synthesized that can self-replicate as well as evolve *in vitro* (153,253). Synthetic polypeptides have also been designed that can self-replicate (254). Surfactant vesicles have been synthesized that can catalyze the formation of more surfactant molecules from their precursors and incorporate them into the growing vesicles or form new vesicles (thus earning the term *self-reproducing vesicles*) (33,255–258). Recently, it has been shown that enzymatic reactions can be carried out within the vesicles to form the constituents of the vesicle itself. The new vesicles share the reactants of the reaction being carried out within them (similar to cell division) (259).

The studies reviewed above demonstrate that chemists have begun to mimic some of the strategies prevalent in biological systems. The design of intelligent substrates, ie, molecular units with all the information necessary for forming supramolecular structures encoded in their primary structure, has already begun. The production of complex, synthetic, functional nanoscale devices by molecular self-assembly of constituent parts is, thus, an important paradigm in the development of nanotechnology.

Conclusions and Outlook

Nobel-laureate Richard Feynman once said that the principles of physics do not preclude the possibility of maneuvering things atom by atom (260). Recent developments in the fields of physics, chemistry, and biology (briefly described in the previous sections) bear those words out. The invention and development of scanning probe microscopy has enabled the isolation and manipulation of individual atoms and molecules. Research in protein and nucleic acid structure have given rise to powerful tools in the establishment of rational synthetic protocols for the production of new medicinal drugs, sensing elements, catalysts, and electronic materials.

Much of the current progress in nanotechnology is confined to understanding and harnessing pre-existing, exquisitely evolved nanomachinery, ie, natural living systems. Many of these developments, eg, biotechnological synthesis of functional proteins (eg, enzymes, antibodies), have progressed to commercial production. Table 2 summarizes the various reductive and synthetic strategies employed to date. The growing scientific and technical progress in nanotechnology will, in the next few decades, continue to produce many new biotechnology-based as well as biomimetic and nonnatural (ie, synthetic) nanodevices. The manufacture of this broad class of products will play a vital role in the world economy of the twenty-first century and the future of humanity.

Table 2. Some Approaches in Reductive (Top Down) and Synthetic (Bottom Up) Nanofabrication

Approach	Examples	References
reductive (also known as top-down)	conventional microfabrication	
	photolithography	16
	x-ray lithography	85
	E-beam lithography	91
	imprint lithography	186,187
	local probe lithography	38,261–263
synthetic (also known as bottom-up)	nanochemical synthesis	5,35,36,145
	biotechnological expression	13,14,264
	molecular templating	178,180–182,188,265
	molecular self-assembly	2,4,17,73
	local probe-assisted synthesis	37,38,56,266,267
hybrid (reductive synthetic)	microcontact printing	183,268
	optical–laser manipulation	102,269–271

ACKNOWLEDGEMENT

Support for this work was provided by ONR Multidisciplinary University Research Initiative Grant N00014 95 1 1315, by ONR Grant N00014 95 1 0901, and by NSF Grant HRD 9450475.

BIBLIOGRAPHY

1.

P. Ball, *Designing the Molecular World: Chemistry at the Frontier*, Princeton University Press, Princeton, 1994.

2.

W. M. Tolles, in G. M. Chow and K. E. Gonsalves, eds., *210th National Meeting of the ACS*, ACS, Chicago, Ill., 1996, pp. 1–15.

3.

J. A. Stroschio and D. M. Eigler, *Science* **254**, 1319–1326 (1991).

4.

G. M. Whitesides and co-workers, *Acc. Chem. Res.* **28**, 37–44 (1995).

5.

G. A. Ozin, *Adv. Mater.* **4**, 612–649 (1992).

6.

D. Philp and J. F. Stoddart, *Angew. Chem. Int. Ed. Engl.* **35**, 1154–1196 (1996).

7.

S. Asai and Y. Wada, *Proc. IEEE* **85**, 505–520 (1997).

8.

K. D. Wise and K. Najfi, *Science* **254**, 1335–1342 (1991).

9.

L. Fabrizzi and A. Poggi, *Chem. Soc. Rev.* **24**, 197–202 (1995).

10.

K. E. Drexler, *Engines of Creation* Anchor Press Doubleday, Garden City, 1986, pp. 1–298.

11.

M. Krummenacker and J. Lewis, *Prospects in Nanotechnology: Toward Molecular Manufacturing*, John Wiley & Sons, Inc., New York, 1995, p. 297.

12.

E. Regis, *Nano: The Emerging Science of Nanotechnology* Little, Brown and Company, Boston, 1995, pp. 1–325.

13.

L. Stryer, *Biochemistry*, W. H. Freeman & Co., New York, 1988.

14.

M. Edelstein, in M. Krummenacker and J. Lewis, eds., *Prospects in Nanotechnology: Toward Molecular Manufacturing*, John Wiley & Sons, Inc., New York, 1995, pp. 67–91.

15.

G. D. Hutcheson and J. D. Hutcheson, *Sci. Am.* **274**, 54–62 (1996).

16.

R. F. Service, *Science* **274**, 1834–1836 (1996).

17.

G. M. Whitesides, J. P. Mathias, and C. T. Seto, *Science* **254**, 1312–1319 (1991).

18.

B. D. Ratner, *J. Biomed. Mater. Res.* **27**, 837–850 (1993).

19.

S. M. Block, *Nature* **386**, 217–219 (1997).

20.

R. R. Birge, *Computer* **25**, 56–67 (1992).

21.

I. Petersen, *Sci. News* **150**, 26 (1996).

22.

J.-M. Lehn, *Angew. Chem. Intl. Ed. Engl.* **29**, 1304–1309 (1990).

23.

P. Hogeweg, *Ber. Bunsenges, Phys. Chem.* **98**, 1135–1139 (1994).

24.

F. Toda, *Bioorg. Chem.* **19**, 157–168 (1991).

25.

K. Johnson, R. K. Allemann, H. Widmer, and S. A. Benner, *Nature* **365**, 530–532 (1993).

26.

S. Borman, *C&EN* **75**, 8–9 (1997).

27.

M. S. Goodman, V. Jubian, B. Linton, and A. D. Hamilton, *J. Am. Chem. Soc.* **117**, 11610–11611 (1995).

28.

E. M. Gordon, M. A. Gallop, and D. V. Patel, *Acc. Chem. Res.* **29**, 144–154 (1996).

29.

D. R. Burton, *Acc. Chem. Res.* **26**, 405–411 (1993).

30.

J. C. Hogan, *Nature Biotech.* **15**, 328–330 (1997).

31.

B. A. Cornell and co-workers, *Nature* **387**, 580–583 (1997).

32.

A. Ulman, *Chem. Rev.* **96**, 1533–1554 (1996).

33.

S. A. Walker, M. Kennedy, and J. A. Zasadzinski, *Biophys. J.* **72**, MP248 (1997).

34.

P. Calvert and P. Rieke, *Chem. Mater.* **8**, 1715–1727 (1996).

35.

R. W. Armstrong and co-workers, *J. Am. Chem. Soc.* **111**, 7525–7530 (1989).

36.

R. B. Woodward, *Pure Appl. Chem.* **33**, 145–177 (1973).

37.

P. Avouris, *Acc. Chem. Res.* **28**, 95–102 (1995).

38.

P. M. Campbell, E. S. Snow, and P. J. McMarr, *Surf. Sci.* **361/362**, 870–873 (1996).

39.

E. S. Snow, P. M. Campbell, and F. K. Perkins, *Proc. IEEE* **85**, 601–611 (1997).

40.

M. Pique, in M. Krummenacker and J. Lewis, eds., *Prospects in Nanotechnology: Toward Molecular Manufacturing*, John Wiley & Sons, Inc., New York, 1995, pp. 93–112.

41.

D. E. Koshland, *Angew. Chem. Int. Ed. Engl.* **33**, 2375–2378 (1994).

42.

C.-J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov, *Crit. Rev. Biochem. Molec. Biol.* **31**, 127–152 (1996).

43.

F. Eisenhaber, B. Persson, and P. Argos, *Crit. Rev. Biochem. Molec. Biol.* **30**, 1–94 (1995).

44.

D. E. Draper, *TIBS* **21**, 145–149 (1996).

45.

I. D. Kuntz, E. C. Meng, and B. K. Shoichet, *Acc. Chem. Res.* **27**, 117–123 (1994).

46.

G. M. Whitesides, *Sci. Am.* **273**, 146–149 (1995).

47.

S. Borman, *C&EN* **74**, 29–35 (1996).

48.

J. H. Weaver, *Naval Res. Rev.* **3**, 16–27 (1991).

49. J. H. Fendler and F. C. Meldrum, *Adv. Mater.* **7**, 607–632 (1995).
50. D. W. Breck, *Zeolite Molecular Sieves*, Kieger, Malabar, 1984.
51. J. Heinze, *Angew. Chem. Int. Ed. Engl.* **30**, 170–171 (1991).
52. D. Mandler, S. Meltzer, and I. Shohat, *Israel J. Chem.* **36**, 73–80 (1996).
53. K. Gamo, *Nuclear Instruments and Methods in Physics Research B* **121**, 464–469 (1997).
54. S. Matsui, *Proc. IEEE* **85**, 629–643 (1997).
55. F. Cerina, *Proc. IEEE* **84**, 644–651 (1997).
56. J. Foster, in B. C. Crandall and J. Lewis, eds., *Nanotechnology: Research and Perspectives*, MIT Press, 1992, pp. 15–36.
57. J. S. Richardson and co-workers, *Biophys. J.* **63**, 1186–1209 (1992).
58. R. J. Ellis and F.-U. Hartl, *FASEB J.* **10**, 20–26 (1996).
59. F. U. Hartl, *Nature* **381**, 571–580 (1996).
60. C. K. Mathews, *J. Bacteriol.* **175**, 6377–6381 (1993).
61. J. J. Kasianowicz, D. L. Burden, L. C. Han, S. Cheley, and H. Bayley, *Biophys. J.* **72**, ThAM-D7 (1997).
62. K. Sigler and M. Hofer, *Crit. Rev. Biotech.* **17**, 69–86 (1997).
63. J. M. Grunkemeier and T. A. Horbett, *J. Molec. Recognition* **9**, 247–257 (1996).
64. J. A. Lopez, *Blood Coagul. Fibrinolysis* **5**, 97–119 (1994).
65. M. J. Russo, H. Bayley, and M. Toner, *Nature Biotech.* **15**, 278–282 (1997).
66. S. Bhaki and co-workers, *Arch. Microbiol.* **165**, 73–79 (1996).
67. D. M. Engelman, *Science* **274**, 1850–1851 (1996).
68. L. Song and co-workers, *Science* **274**, 1859–1866 (1996).
69. R. G. Panchal and H. Bayley, *J. Biol. Chem.* **270**, 23072–23076 (1995).
70. C. Y. Chang, B. Niblack, B. Walker, and H. Bayley, *Chem. and Biol.* **2**, 391–400 (1995).
71. R. M. Ghadiri, J. R. Granja, and L. K. Buehler, *Nature* **369**, 301–304 (1994).
72. M. R. Ghadiri, J. R. Granja, R. A. Milligan, D. E. McRee, and N. Khazanovich, *Nature* **366**, 324–327 (1993).
73. M. R. Ghadiri, *Adv. Mater.* **7**, 675–677 (1995).
74. M. R. Ghadiri, K. Kobayashi, J. R. Granja, R. K. Chada, and D. E. McRee, *Angew. Chem. Int. Ed. Engl.* **34**, 93–95 (1995).
75. B. J. Schnapp, *Nature* **373**, 655 (1995).
76. S. M. Block, *Cell* **87**, 151–157 (1996).
77. L. G. Tilney and D. A. Portnoy, *J. Cell Biol.* **109**, 1597–1608 (1989).
78. K. Ogawa and H. Mohri, *Cell Struct. Funct.* **21**, 343–349 (1996).
79. S. C. West, *Cell* **86**, 177–180 (1996).
80. A. D. T. Samuel and H. G. Berg, *Biophys. J.* **71**, 918–923 (1996).
81. T. Thompson, *Byte* **21**, 45–54 (1996).
82. A. L. Robinson, *Science* **223**, 267–268 (1984).
83. D. Goldhaber-Gordon, M. S. Montemero, J. C. Love, J. O. Gregory, and J. C. Ellenbogen, *Proc. IEEE* **85**, 521–540 (1997).
84. Y. Taur and co-workers, *Proc. IEEE* **85**, 486–504 (1997).
85. H. I. Smith, *Phys. Scripta* **T61**, 26–31 (1996).
86. A. Snigirev, V. Kohn, I. Snigireva, and B. Lengeler, *Nature* **384**, 49–51 (1996).
87. S. H. Zaidi and S. R. J. Brueck, *J. Vac. Sci. Technol. B* **11**, 658–666 (1993).
88. S. H. Zaidi and S. R. J. Brueck, in T. A. Brunner, ed., *Proc. SPIE: Optical/Laser Microolithography VII*, **2197**, San Jose, CA, 869–875, 1994.
89. S. Chu, *S. Sci. Am.*, 49–54 (1992).
90. K. K. Berggren and co-workers, *Science* **269**, 1255–1257 (1995).
91. L. Harriott and A. Liddle, *Physics World* **10**, 41–45 (1997).
92. R. R. Birge, *Optics Letters* **20**, 2429–2431 (1995).
93. J. W. Lyding, *Proc. IEEE* **85**, 589–600 (1997).
94. M. F. Crommie, C. P. Lutz, and D. M. Eigler, *Science* **262**, 218 (1993).
95. K. Bessho, Y. Iwasaki, and S. Hashimoto, *IEEE Trans. Magnetis* **32**, 4443–4447 (1996).
96. M. Jaschke and H.-J. Butt, *Langmuir* **11**, 1061–1064 (1995).
97. D. Keller, *Nature* **384**, 111 (1996).
98. W. Xie, X. Dai, L. S. Xu, D. A. Allee, and J. Spector, *Nanotechnology* **8**, 88–93 (1997).
99. D. M. Kolb, R. Ullmann, and T. Will, *Science* **275**, 1097–1099 (1997).
100. P. D. Lett and co-workers, *J. Opt. Soc. Am. B* **6**, 2084–2107 (1989).
101. P. Brumer and M. Shapiro, *Sci. Am.*, 56–63 (1995).
102. S. Sato and H. Inaba, *Optical and Quantum Electronics* **28**, 1–16 (1996).
103. M. R. Emmert-Buck and co-workers, *Science* **274**, 998–1001 (1996).
104. H. P. Erickson, *Science* **276**, 1090–1092 (1997).
105. M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier, and C. Bustamante, *Science* **276**, 1112–1116 (1997).
106. L. Tskhovrebova, J. Trinick, J. A. Sleep, and R. M. Simmons, *Nature* **387**, 308–312 (1997).
107. R. G. Larson, T. T. Perkins, D. E. Smith, and S. Chu, *Phys. Rev. E* **55**, 1794–1797 (1997).
108. K. Helmersen, R. Kishore, W. D. Phillips, and H. H. Weetall, *Clin. Chem.* **43**, 379–383 (1997).
109. D. E. Leckband, F. J. Schmitt, J. N. Israelachvili, and W. Knoll, *Biochem.* **33**, 4611–4624 (1994).
110. C. L. Bowes and G. A. Ozin, *Adv. Mater.* **8**, 13–28 (1996).
111. G. A. Ozin and S. Oliver, *Adv. Mater.* **7**, 943–951 (1995).
112. M. D. Ward, in B. C. Crandall and J. Lewis, eds., *Nanotechnology: Research and Perspectives*, MIT Press, 1992, pp. 67–103.
113. G. Georgiou and co-workers, *Nature Biotech.* **15**, 29–34 (1997).
114. F. Michel and E. Westhof, *Science* **273**, 1676–1677 (1996).
115. D. W. Celander and T. R. Cech, *Science* **251**, 401–407 (1991).
116. K. J. Limbach and E. Paoletti, *Epidemiol. Infect.* **116**, 241–256 (1996).
117. J. R. Lorsch and J. W. Szostak, *Acc. Chem. Res.* **29**, 103–110 (1996).
118. R. B. Merrifield, *J. Am. Chem. Soc.* **85**, 2149–2154 (1963).
119. P. Buckel, *Trends in Pharmacological Sciences* **17**, 450–456 (1996).
120. R. D. Palmiter, G. Norstedt, R. E. Gelinas, R. E. Hammer, and R. L. Brinster, *Science* **222**, 809–814 (1983).
121. J. C. Erickson, G. Hollopeter, and R. D. Palmiter, *Science* **274**, 1704–1707 (1996).
122. F. A. Fletcher and co-workers, *Experimental Hematology* **24**, 172 (1996).
123. A. S. Moffat, *Science* **275**, 757 (1997).

124. D. Cai and co-workers, *Science* **275**, 832–834 (1997).

125. F. R. Haselton, W. C. Tseng, and M. S. Chang, *Investigative Ophthalmology and Visual Science* **38**, 2082 (1997).

126. M. S. Chang and F. R. Haselton, *Investigative Ophthalmology and Visual Science* **38**, 2083 (1997).

127. A. R. Thierry, P. Rabinovich, B. Peng, L. C. Mahan, J. L. Bryant, and R. C. Gallo, *Gene Therapy* **4**, 226–237 (1997).

128. L. Huang and S. Li, *Nature Biotech.* **15**, 620–621 (1997).

129. V. D'Hondt, M. Caruso, and A. Bank, *Blood* **88**, 1711 (1996).

130. A. Rovira, A. Wong, M. Sadelain, and L. Luzzatto, *Blood* **88**, 3902 (1996).

131. J. Light, R. Maki, and N. Assa-Munt, *Nucleic Acids Research* **25**, 4367–4368 (1996).

132. V. D'Hondt, M. Caruso, M. Ward, and A. Bank, *Brit. J. Haematology* **93**, 1266 (1996).

133. J. Wohlgenuth, S. H. Kang, G. H. Bulboaca, K. Nawotka, and M. P. Calos, *Gene Therapy* **3**, 503–512 (1996).

134. A. Bielinska, J. F. Kukowska-Latallo, J. Johnson, D. A. Tomalia, and J. R. Baker, *FASEB J.* **10**, 1109 (1996).

135. A. Baulard and co-workers, *Gene* **176**, 149–154 (1996).

136. M. Onishi and co-workers, *Experimental Hematology* **24**, 324–329 (1995).

137. J. Roth and co-workers, *Proc. Natl. Acad. Sci. USA* **93**, 4781–4786 (1996).

138. V. Zhang, M. O'Sullivan, H. Hamed, W. T. Roswit, and M. J. Holtzman, *Biochem. Biophys. Res. Comm.* **227**, 499–506 (1996).

139. I. Frolov and co-workers, *Proc. Natl. Acad. Sci. USA* **93**, 11371–11377 (1996).

140. J. P. Levy, R. R. Muldoon, S. Zolotukhin, and C. J. Link, *Nature Biotech.* **14**, 610–614 (1996).

141. A. M. Hunter, E. T. Rhodes, and J. C. Lane, *Molec. Biol. Cell* **7**, 3871 (1996).

142. Y. Dai, E. M. Schwartz, D. Gu, W. W. Zhang, N. Sarvetnick, and I. M. Verma, *Scientist* **11**, 16 (1997).

143. A. Eschenmoser and C. E. Wintner, *Science* **196**, 1410–1420 (1977).

144. P. P. Deshpande and S. J. Danishefsky, *Nature* **387**, 164–166 (1997).

145. K. C. Nicolaou and co-workers, *Nature* **387**, 268–272 (1997).

146. N. J. Turro, J. K. Barton, and D. A. Tomalia, *Acc. Chem. Res.* **24**, 332–340 (1991).

147. D. S. Lawrence, T. Jiang, and M. Levett, *Chem. Rev.* **95**, 2229–2260 (1995).

148. W. Saenger, *Angew. Chem. Int. Ed. Engl.* **19**, 344–362 (1980).

149. H. Ogino, *J. Am. Chem. Soc.* **103**, 1303–1304 (1981).

150. H. Ogino and K. Ohata, *Inorganic Chemistry* **23**, 3312–3316 (1984).

151. G. Lowe, *Chem. Rev.* 309–317 (1996).

152. K. D. Janada and co-workers, *Science* **275**, 945–948 (1997).

153. R. Rawls, *C&EN*, 10, April 28 (1997).

154. J. A. Ellman, *Acc. Chem. Res.* **29**, 132–143 (1996).

155. R. Breslow, *Acc. Chem. Res.* **28**, 146–153 (1995).

156. K. S. Lam and co-workers, *Nature* **354**, 82–86 (1991).

157. W. C. Still, *Acc. Chem. Res.* **29**, 155–163 (1996).

158. X.-D. Xiang and co-workers, *Science* **268**, 1738–1740 (1995).

159. N. Williams, *Science* **277**, 474–477 (1997).

160. R. Taylor and D. R. M. Walton, *Nature* **363**, 685–693 (1993).

161. H. Terrones, M. Terrones, and W. K. Hsu, *Chem. Rev.*, 341–350 (1996).

162. P. Ball, *New Scientist*, 28–31 (1996).

163. C. A. Mirkin and W. B. Calwell, *Tetrahedron* **52**, 5113–5130 (1996).

164. W. Z. Li and co-workers, *Science* **274**, 1701–1703 (1996).

165. A. Thess and co-workers, *Science* **273**, 483–487 (1996).

166. S. J. Tans and co-workers, *Nature* **386**, 474–477 (1997).

167. R. S. Lee, H. J. Kim, J. E. Fischer, A. Thess, and R. E. Smalley, *Nature* **388**, 255–257 (1997).

168. M. Bockrath and co-workers, *Science* **275**, 1922–1925 (1997).

169. A. C. Dillon and co-workers, *Nature* **386**, 377–379 (1997).

170. D. Ugarte, A. Chatelain, and W. A. De Heer, *Science* **274**, 1897–1899 (1996).

171. G. E. Gadd and co-workers, *Science* **277**, 933–936 (1997).

172. H. Dai, J. H. Hafner, A. G. Rinzler, D. T. Colbert, and R. E. Smalley, *Nature* **384**, 147–150 (1996).

173. C. T. Kresge, *Adv. Mater.* **8**, 181–182 (1996).

174. N. K. Raman, M. T. Anderson, and C. J. Brinker, *Chem. Mater.* **8**, 1682–1701 (1996).

175. C. T. Kresge, M. E. Leonowicz, W. J. Roth, J. C. Vartuli, and J. S. Beck, *Nature* **359**, 710–712 (1992).

176. G. Cao, Y. Lu, L. Delattre, C. J. Brinker, and G. P. Lopez, *Adv. Mater.* **8**, 588–591 (1996).

177. C. J. Brinker and G. W. Scherer, *Sol-Gel-Science: The Physics and Chemistry of Sol-Gel Processing*, Academic Press, San Diego, Calif., 1990.

178. R. Hoss and F. Vogtle, *Angew. Chem. Int. Ed. Engl.* **33**, 375–384 (1994).

179. Z. Cai and C. R. Martin, *J. Am. Chem. Soc.* **111**, 4138–4139 (1989).

180. C. R. Martin, L. S. Van Dyke, Z. Cai, and W. Liang, *J. Am. Chem. Soc.* **112**, 8976–8977 (1990).

181. C. R. Martin, *Acc. Chem. Res.* **28**, 61–68 (1995).

182. Y. Xia and co-workers, **273**, 347–349 (1996).

183. Y. Xia and co-workers, *Adv. Mater.* **9**, 147–149 (1997).

184. Y. Xia, D. Qin, and G. M. Whitesides, *Adv. Mater.* **8**, 1015–1017 (1996).

185. X.-M. Zhao and co-workers, *Adv. Mater.* **8**, 420–424 (1996).

186. S. Y. Chou, P. R. Krauss, and P. J. Renstrom, *Science* **272**, 85–87 (1996).

187. S. Y. Chou, P. R. Krauss, and P. J. Renstrom, *J. Vac. Sci. Technol. B* **14**, 4129–4133 (1996).

188. K. Mosbach and O. Ramstrom, *Bio/Technology* **14** 163–170, (1996).

189. R. H. Tredgold, *J. Mater. Chem.* **5**, 1095–1106 (1995).

190. G. J. Ashwell, E. J. C. Dawnay, A. P. Kuzynski, and P. J. Martin, *Proc. SPIE, Int. Soc. Eng.* **1361**, 589 (1991).

191. W. H. A. Majid, T. Richardson, S. Holder, and D. Lacey, *Thin Solid Films* **243**, 378–383 (1994).

192. M. R. Bryce and M. C. Petty, *Nature* **374**, 771–776 (1995).

193. D. L. Allara, *Biosensors Bioelectronics* **10**, 771–783 (1995).

194. L. H. Dubois and R. G. Nuzzo, *Ann. Rev. Phys. Chem.* **43**, 437–463 (1992).

195. A. Ulman, *Chem. Rev.* **96**, 1533–1544 (1996).

196. J. L. Wilbur, A. Kumar, E. Kim, and G. M. Whitesides, *Adv. Mater.* **6**, 600–604 (1994).

197. J. L. Wilbur, R. J. Jackman, G. M. Whitesides, E. L. Cheung, L. K. Lee, and M. G. Prentiss, *J. Am. Chem. Soc.* **8**(3), 825–831 (1995).

198. M. Mrksich, J. R. Grunwell, and G. M. Whitesides, *J. Am. Chem. Soc.* **117**(48), 12009–12010 (1995).

199. S. Spinke, M. Liley, H.-J. Guder, L. Angermaier, and W. Knoll, *Langmuir* **9** (1993).

200. L. M. Tender, R. L. Worley, H. Fan, and G. P. Lopez, *Langmuir* **12**, 5515–5518 (1996).

201. L. Sun and R. M. Crooks, *Langmuir* **9**, 1775–1780 (1993).

202. C. D. Bain, Y. T. Tao, J. Eval, G. M. Whitesides, and R. G. Nuzzo, *J. Am. Chem. Soc.* **111**, 321–335 (1989).

203. P. E. Laibinis and co-workers, *J. Am. Chem. Soc.* **113**, 7152–7167 (1991).

204. L. Strong and G. M. Whitesides, *Langmuir* **4**, 546–558 (1988).

205. P. Fenter, A. Eberhardt, and P. Eisenberger, *Science* **266**, 1216–1218 (1994).

206. P. A. Di Milla and co-workers, *J. Am. Chem. Soc.* **116**, 2225–2226 (1994).

207. P. E. Laibinis, R. G. Nuzzo and G. M. Whitesides, *J. Phys. Chem.* **96**, 5097–5105 (1992).

208. R. Singhvi and co-workers, *Science* **264**, 696–698 (1994).

209. G. P. Lopez, H. A. Biebuyck, R. Haerter, A. Kumar, and G. M. Whitesides, *J. Am. Chem. Soc.* **115**, 10774–10781 (1993).

210. G. P. Lopez and G. M. Whitesides, *Science* **260**, 647–649 (1993).

211. N. A. Abbott and co-workers, *NATO ASI Ser., Ser. E* **239**, 293–301 (1993).

212. C. B. Ross, L. Sun, and R. M. Crooks, University of New Mexico, 1992.

213. R. C. Tiberio and co-workers, *Appl. Phys. Lett.* **62**, 476–478 (1993).

214. C. D. Bain and co-workers, *J. Am. Chem. Soc.* **111**, 321–335 (1989).

215. N. L. Abbott and G. M. Whitesides, *Langmuir* **10**, 1493–1497 (1994).

216. K. R. Stewart and G. M. Whitesides, *Rev. Sci. Instrum.* **57**, 1381–1383 (1986).

217. M. Garcia-Parajo, C. Longo, J. Servant, P. Gorostiza, and F. Sanz, *Langmuir* **13**, 2333–2339 (1997).

218. P. E. Laibinis and G. M. Whitesides, *J. Am. Chem. Soc.* **114**, 9022–9028 (1992).

219. K. L. Prime and G. M. Whitesides, *J. Am. Chem. Soc.* **115**, 10714–10721 (1993).

220. G. P. Lopez and co-workers, *J. Am. Chem. Soc.* **115**, 5877–5878 (1993).

221. J. M. Calvert, *J. Vac. Sci. Technol. B* **11**, 2155–2163 (1993).

222. L. Netzer and J. Sagiv, *J. Am. Chem. Soc.* **105**, 674–676 (1983).

223. Y. W. Lee, J. Reed-Mundle, C. N. Sukenik, and J. E. Zull, *Langmuir* **9**, 3009–3014 (1993).

224. S. I. Stupp and co-workers, *Science* **276**, 384–389 (1997).

225. D. J. Cram, *Science* **240**, 760–767 (1988).

226. D. J. Cram, *Nature* **356**, 29–36 (1992).

227. T. E. Mallouk, *Nature* **387**, 350–351 (1997).

228. G. S. Hanan, C. R. Arana, J.-M. Lehn and D. Fenske, *Angew. Chem. Int. Engl.* **34**, 1122–1124 (1995).

229. P. N. W. Baxter, J.-M. Lehn, J. Fischer, and M.-T. Youinou, *Angew. Chem. Int. Ed. Engl.* **33**, 2284–2287 (1994).

230. H. Tsukube, *Coord. Chem. Rev.* **148**, 1–17 (1996).

231. M. Fujita, J. Yazaki, and K. Ogurqa, *J. Am. Chem. Soc.* **112**, 5645–5647 (1990).

232. M. Fujita, J. Yazaki, and K. Ogura, *Chem. Lett.*, 1031–1032 (1991).

233. Z. Li, C. M. Wang, L. Persaud, and T. E. Mallouk, *J. Phys. Chem.* **92**, 2592–2597 (1988).

234. L. Persaud and co-workers, *J. Am. Chem. Soc.* **109**, 7309–7314 (1987).

235. Z. Li, C. Lai, and T. E. Mallouk, *Inorg. Chem.* **28**, 178–182 (1989).

236. L. Zelikovich, J. Libman, and A. Shanzer, *Nature* **374**, 790–792 (1995).

237. J. P. Mathias, C. T. Seto, E. E. Simanek, and G. M. Whitesides, *J. Am. Chem. Soc.* **116**, 1725–1736 (1994).

238. J. P. Mathias, E. E. Simanek, C. T. Seto, and G. M. Whitesides, *Angew. Chem. Int. Ed. Engl.* **32**, 1766–1769 (1993).

239. J. F. Stoddart, *Carbohydrate Research* **192**, xii–xv (1989).

240. N. Khazanovich, J. R. Granja, D. E. McRee, R. A. Milligan, and M. R. Ghadiri, *J. Am. Chem. Soc.* **116**, 6011–6012 (1994).

241. M. R. Ghadiri, C. Soares, and C. Choi, *Angew. Chem. Int. Ed. Eng.* **32**, 1594–1597 (1992).

242. M. R. Ghadiri, C. Soares, and C. Choi, *J. Am. Chem. Soc.* **114**, 825–831 (1992).

243. M. R. Ghadiri, C. Soares, and C. Choi, *J. Am. Chem. Soc.* **114**, 4000–4002 (1992).

244. M. Griffith, B. Oskouian, and M. R. Ghadiri, *FASEB J.* **7**, A1275 (1993).

245. N. C. Seeman, Y. Zhang, and J. Chen, *J. Vac. Sci. Technol. A* **12**, 1895–1902 (1994).

246. E. A. Wintner, M. M. Conn, and J. Rebek, Jr., *Acc. Chem. Res.* **27**, 198–203 (1994).

247. R. J. Pieters, I. Huc, and J. Rebek, Jr., *Angew. Chem. Intl. Ed. Engl.* **33**, 1579–1581 (1994).

248. M. M. Conn, E. A. Wintner, and J. Rebek, Jr., *Angew. Chem. Intl. Ed. Engl.* **33**, 1577–1579 (1994).

249. T. Tjivikau, P. Ballester, and J. Rebek, *J. Am. Chem. Soc.* **112**, 1249–1250 (1990).

250. J. S. Nowick, Q. Feng, T. Tjivikua, P. Ballester, and J. Rebek, *J. Am. Chem. Soc.* **113**, 8831–8839 (1991).

251. G. F. Joyce, *Science* **276**, 1658–1659 (1997).

252. S. Chiruvolus and co-workers, *Science* **264**, 1753–1756 (1994).

253. M. C. Wright and G. F. Joyce, *Science* **276**, 614–617 (1997).

254. D. H. Lee, J. R. Granja, J. A. Martinez, K. Severin, and M. R. Ghadiri, *Nature* **382**, 525–528 (1996).

255. R. Maoz, S. Matlism, E. DiMasi, B. M. Ocko, and J. Sagiv, *Nature* **384**, 150–153 (1996).

256. J. E. Rothman and L. Orci, *Sci. Am.* **273**, 70–75 (1996).

257. S. Keller and co-workers, *Biophys. J.* **72**, TH307 (1997).

258. S. A. Walker, M. T. Kennedy, and J. A. Zasadzinski, *Nature* **387**, 61–64 (1997).

259. R. Wick and P. L. Luisi, *Chem. Biol.* **3**, 277–285 (1996).

260. R. P. Feynman, *Eng. Sci.* **23**, 22–36 (1960).

261. J. W. Lyding and co-workers, *Israel J. Chem.* **36**, 3–10 (1996).

262. H. M. Marchman, J. E. Griffith, J. Z. Y. Guo, J. Frackoviak, and G. K. Celler, *J. Vac. Sci. Technol. B* **12**, 3585–3590 (1994).

263. F. Cerrina and C. Marrian, *MRS Bulletin*, 56–62 (1996).

264. S. M. Edgington, *Bio/Technol.* **12**, 468–471 (1994).

265. G. Bar, S. Rubin, R. W. Cutts, T. N. Taylor, and T. A. Zawodzinski, Jr., *Langmuir* **12**, 1172–1179 (1996).

266. J. Jersch, F. Demming, and K. Dickman, *Applied Physics A.* **64**, 29–32 (1997).

267. K. Douglas, N. A. Clark, and K. J. Rothschild, *Applied Physics Letters* **48**, 676–678 (1986).

268. A. Kumar, N. L. Abbott, E. Kim, H. A. Biebuyck, and G. M. Whitesides, *Acc. Chem. Res.* **28**, 219–226 (1995).

269. L. J. Kricka, *Clinical Chemistry* **43**, 251–253 (1997).

270. R. E. Behringer, V. Natarajan, and G. Timp, *Appl. Phys. Lett.* **68**, 1034–1036 (1996).

271. W. D. Phillips and co-workers, *IEEE Trans. Instrum. Measur.* **40**, 78–80 (1991).

Rajesh Vaidya
Gabriel Lypez
University of New Mexico

Jos   A. Lypez
Baylor College of Medicine

NITRILES

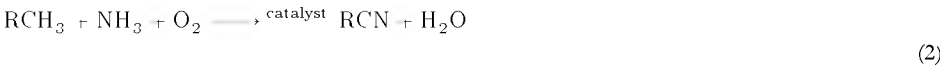
Nitriles, or organic cyanides, are organic compounds which contain the cyano (ie, -CN) group. Nitriles are often considered derivatives of carboxylic acids and are named according to the carboxylic acid which is produced upon hydrolysis of the nitrile. For example, cyanomethane (methyl cyanide) is named acetonitrile [75-05-8] because hydrolysis of its cyano group yields acetic acid. Nitriles which contain additional functional groups are typically named as cyano-substituted compounds (eg, cyanoacetic acid). Nitriles which contain a hydroxy (–OH) group on the carbon atom that is bonded to the cyano moiety are known as cyanohydrins (qv). According to *Chemical Abstracts*, aliphatic nitriles are named as derivatives of the longest carbon chain and the carbon of the nitrile is included.

General Preparations and Chemical Properties

Nitriles may be prepared by several methods (1). The first nitrile to be prepared was propionitrile, which was obtained in 1834 by distilling barium ethyl sulfate with potassium cyanide. This is a general preparation of nitriles from sulfonate salts and is referred to as the Pelouze reaction (2). Although not commonly practiced today, dehydration of amides has been widely used to produce nitriles and was the first commercial synthesis of a nitrile. The reaction of alkyl halides with sodium cyanide to produce nitriles (eq. 1) also is a general reaction with wide applicability:



where X = Cl, Br, or I. If dimethyl sulfoxide is used as solvent, high yields of nitriles can be obtained with both primary and secondary alkyl chlorides (see SULFOXIDES). This method also may be used for preparing dinitriles. Reaction times usually are less than one hour (3).
Ammonoxidation, a vapor-phase reaction of hydrocarbon with ammonia and oxygen (air) (eq. 2), can be used to produce hydrogen cyanide (HCN), acrylonitrile, acetonitrile (as a by-product of acrylonitrile manufacture), methacrylonitrile, benzonitrile, and toluinitriles from methane, propylene, butylene, toluene, and xylenes, respectively (4).



Processes have been developed whereby the oxygen is supplied from the crystal lattice of a metal-oxide catalyst (5) (see ACRYLONITRILE; METHACRYLIC ACID AND DERIVATIVES).
Addition of HCN to unsaturated compounds is often the easiest and most economical method of making organonitriles. An early synthesis of acrylonitrile involved the addition of HCN to acetylene. The addition of HCN to aldehydes and ketones is readily accomplished with simple base catalysis, as is the addition of HCN to activated olefins (Michael addition). However, the addition of HCN to unactivated olefins and the regioselective addition to dienes is best accomplished with a transition-metal catalyst, as illustrated by DuPont's adiponitrile process (6–9).

Chemistry and Uses of Nitriles

As a class of compounds, nitriles have broad commercial utility that includes their use as solvents, feedstocks, pharmaceuticals, catalysts, and pesticides. The versatile reactivity of organonitriles arises both from the reactivity of the C–N bond, and from the ability of the cyano substituent to activate adjacent bonds, especially C–H bonds. Nitriles can be used to prepare amines, amides, amidines, carboxylic acids and esters, aldehydes, ketones, large-ring cyclic ketones, imines, heterocycles, orthoesters, and other compounds. Some of the more common transformations involve hydrolysis or alcoholysis to produce amides, acids and esters, and hydrogenation to produce amines, which are intermediates for the production of polyurethanes and polyamides. An extensive review on hydrogenation of nitriles has been recently published (10).
Acrylonitrile [107-13-1] (qv), the largest volume organonitrile, is an important monomer both for plastics and synthetic fibers. Acetonitrile, a by-product of acrylonitrile manufacture, is commercially important for solvent extraction, reaction media, and as an intermediate in the preparation of pharmaceuticals (qv) and other organic chemicals (see EXTRACTION, LIQUID LIQUID EXTRACTION). Propionitrile [107-12-0], a by-product of the electrodimerization of acrylonitrile to adiponitrile, is used as a chemical intermediate. Hydrogenation of organonitriles to amines provides important intermediates both for polyurethanes (by way of isocyanates) and polyamides (nylons); adiponitrile is used almost exclusively by the manufacturers in the production of 1,6-diaminohexane (hexamethylenediamine), an intermediate for nylon-6,6. Other nitriles that are produced in thousands of metric tons per year include acetone cyanohydrin, 2-amino-2-methylpropionitrile, and fatty acid nitriles. Acetone cyanohydrin is an intermediate for the preparation of methyl methacrylate and acrylic resins, (eg, lucite and plexiglas) and for 5,5-dimethylhydantoin, which is used to make commercial water treatment chemicals. 2-Amino-2-methylpropionitrile is an intermediate for the preparation of azobis(isobutyronitrile), which is a widely used polymerization initiator (eg, Vazo 64) and in the production of some agrichemicals. Other aminonitriles are unisolated intermediates in the production of chelants such as ethylenediaminetetraacetate (EDTA) and nitrilotriacetate (NTA). The fatty acid nitriles are intermediates in the production of a large variety of commercial amines and amides. The nitriles that are commercially available in the United States (1994) and their manufacturers are listed in Table 1 (11). The U.S. Department of Commerce estimates that more than 2.5 million metric tons of nitrile compounds were produced in the United States in 1994. Global production of nitriles has increased significantly in the last twenty years. Excellent resources for identifying producers of organonitriles worldwide include Chem Sources International and SRI International's *Directory of Chemical Producers*.

Table 1. Commercially Available Nitriles in the United States, 1994^a

Nitriles	CAS Registry Number	Manufacturers
acetone cyanohydrin	[75-86-5]	BP Chemicals, CYRO, ICI, Rohn & Hass
acetonitrile	[75-05-8]	BP Chemicals, Cytec, DuPont, JT Baker, Sterling
acrylonitrile	[107-13-1]	BP Chemicals, Cytec, DuPont, Monsanto, Sterling, Witco
adiponitrile	[111-69-3]	DuPont, Monsanto
2-amino-2-methylpropanenitrile	[19355-69-2]	Degussa, DuPont
2-amino-2-methylbutanenitrile	[4475-95-0]	DuPont
2-amino-2,4-dimethylpentane-nitrile	[26842-43-3]	DuPont
1-aminocyclohexanecarbonitrile	[5496-10-6]	DuPont
2,2'-azobis(2-methylpropane-nitrile)	[78-67-1]	DuPont, Wako

2,2′azobis(2-methylbutanenitrile)	[13472-08-7]	DuPont, Wako
2,2′-azobis(dimethylpentanenitrile)	[4419-11-8]	DuPont, Wako
2,2′-azobis(4-methoxy-2,4-dimethylpentanenitrile)	[15545-97-8]	Wako
1,1′-azobis(cyanocyclohexane)	[2094-98-6]	DuPont
benzonitrile	[100-47-0]	PMC
<i>n</i> -butyronitrile	[109-74-0]	Eastman Chemicals
<i>i</i> -butyronitrile	[78-82-0]	Eastman Chemicals
cyanoacetic acid	[372-09-8]	Huls America
-methyl ester	[105-34-0]	Huls America
-ethyl ester	[105-56-6]	Huls America
isophthalonitrile	[626-17-5]	DuPont, Twin Lakes
2-methyl-3-butenenitrile	[16529-56-9]	DuPont
2-methylglutaronitrile	[4553-62-2]	DuPont
2-pentenenitrile	[13284-42-9]	DuPont
3-pentenenitrile	[4635-87-4]	DuPont
propionitrile	[107-12-0]	Monsanto
miscellaneous alkyl and fatty acid nitriles		Akzo Nobel, Dixie, Eneco, Hampshire, Henkel, High Point, Hilton Davis, Olin, Witco ^b

^a Ref. 11.
^b *Bought by GNI in 1995.

General Health and Safety Factors

As a class of compounds, the two main toxicity concerns for nitriles are acute lethality and osteolathyrsm. A comprehensive review of the toxicity of nitriles, including detailed discussion of biochemical mechanisms of toxicity and structure-activity relationships, is available (12). Nitriles vary broadly in their ability to cause acute lethality and subtle differences in structure can greatly affect toxic potency. The biochemical basis of their acute toxicity is related to their metabolism in the body. Following exposure and absorption, nitriles are metabolized by cytochrome p450 enzymes in the liver. The metabolism involves initial hydrogen abstraction resulting in the formation of a carbon radical, followed by hydroxylation of the carbon radical. Metabolism at the carbon atom adjacent (alpha) to the cyano group would yield a cyanohydrin metabolite, which decomposes readily in the body to produce cyanide. Hydroxylation at other carbon positions in the nitrile does not result in cyanide release.

The propensity of nitriles to release cyanide subsequent to metabolism is the basis of their acute toxicity. Nitriles that form tertiary radicals at their alpha carbon atoms (eg, isobutyronitrile, 2-methylbutyronitrile) are substantially more acutely lethal than nitriles that form secondary radicals at their alpha carbons (eg, butyronitrile, propionitrile). Cyanohydrins are acutely toxic because they are unstable and release cyanide quickly. Alpha-aminonitriles are also acutely toxic, presumably by analogy with cyanohydrins.

Osteolathyrism is a disease characterized by lameness, skeletal deformities, aortic aneurysms, and slowing or cessation of body growth. Certain aminonitriles are known to cause osteolathyrism in several animal species; aminoacetoneitrile, 3-aminopropionitrile, and 3-amino-2-methylpropionitrile are potent inducers of this disease. It is believed that the basis of their ability to cause osteolathyrism is due to their ability to inhibit lysine oxidase, an enzyme important in cross-linking of collagen and formation of connective tissue. While no reports of these substances causing osteolathyrism in humans have appeared, it seems plausible that these substances could cause this disease in humans because they are known to cause it in a variety of experimental animals.

Persons handling nitriles should take precautions to prevent inhalation of fumes or skin contact. Butyl rubber gloves are frequently used because they resist permeation by many nitriles more efficiently than most gloves. Inhalation may cause headache, nausea, vomiting, irritation of mucosal membranes, and dizziness. Skin contact may result in similar symptoms, but may develop more slowly. Other symptoms of cyanide poisoning include rapid pulse, flushing of the skin, and initial rapid respiration followed by labored breathing. A person who has inhaled the vapors should be moved to an uncontaminated environment, and cardiopulmonary resuscitation should be administered if necessary. If the victim breathes with difficulty, oxygen should be given. In case of eye contact flush with copious water for at least 20 min and call a physician. In case of ingestion, induce vomiting and call a physician. For skin contact, wash with plenty of soap and water. In each case, observe victim for several hours for delayed reaction due to metabolism or decomposition in the body.

Acetonitrile

Acetonitrile (ethanenitrile), CH₃CN, is a colorless liquid with a sweet, ethereal odor. Its physical properties are listed in Table 2. It is completely miscible with water and its high dielectric strength and dipole moment make it an excellent solvent for both inorganic and organic compounds including polymers. Some representative inorganic compounds that are soluble in acetonitrile are CaCl₂, CuCl, FeCl₂, FeCl₃, KSCN, KMnO₄, AgNO₃, and ZnCl₂. Many gases also are highly soluble in acetonitrile, eg, olefinic hydrocarbons, halides, HCl, SO₂, and H₂S. It forms low boiling azeotropes with many organics and high boiling azeotropes with BF₃, SiCl₄, and (CH₃)₄Pb (13). Some of its azeotropes are listed in Table 3 (14).

Table 2. Some Physical Properties of Acetonitrile

Property	Value
CAS Registry Number	[75-05-8]
mol wt	41.05
bp (at 101.3 kPa = 1 atm), °C	81.6
freezing pt, °C	45.7
density (at 20°C), g cm ³	0.786
<i>n</i> _D ²⁰	1.3441
viscosity (at 20°C), mPa·s (cP)	0.35
heat of vaporization (at 80°C), J kg ^a	72.7 × 10 ⁴
heat of fusion (at 45.7°C), J kg ^a	21.8 × 10 ⁴
heat of combustion (at 25°C), J kg ^a	31.03 × 10 ⁶
heat capacity (liquid at 20°C), J (kg·K) ^a	22.59 × 10 ²
surface tension, mN m (dyn cm)	29.3
coefficient of expansion (at 20°C per °C)	1.37 × 10 ⁻³
specific conductance (at 25°C), S	(5–9) × 10 ⁻⁸
dipole moment, Cm ^b	10.675 × 10 ⁻³⁰

dielectric constant, °C	
at 0	42.0
at 20	38.8
at 81.6	26.2
evaporation rate (butyl acetate = 100)	579
flash pt (COC ^c), °C	6
flammable limits (in air), vol %	
lower	4.4
upper	16.0

^a To convert J to cal, divide by 4.184.
^b To convert Cm to D, divide by 3.336 × 10^{−30}.
^c COC — Cleveland open cup.

Table 3. Acetonitrile Binary Azeotropes^a

Component	Bp (at 101.3 kPa = 1 atm)	Acetonitrile, wt %
benzene	73	34
carbon tetrachoride	65	17
1,2-dichloroethane	79	79
ethanol	73	44
ethyl acetate	75	23
methanol	63	81
water	77	84

^a Ref. 14.

Although acetonitrile is one of the more stable nitriles, it undergoes typical nitrile reactions and is used to produce many types of nitrogen-containing compounds, eg, amides (15), amines (16,17); higher molecular weight mono- and dinitriles (18,19); halogenated nitriles (20); ketones (21); isocyanates (22); heterocycles, eg, pyridines (23), and imidazolines (24). It can be trimerized to *s*-trimethyltriazine (25) and has been telomerized with ethylene (26) and copolymerized with α-epoxides (27).

Most, if not all, of the acetonitrile that was produced commercially in the United States in 1995 was isolated as a by-product from the manufacture of acrylonitrile by propylene ammoxidation. The amount of acetonitrile produced in an acrylonitrile plant depends on the ammoxidation catalyst that is used, but the ratio of acetonitrile:acrylonitrile usually is ca 2–3:100. The acetonitrile is recovered as the water azeotrope, dried, and purified by distillation (28). U.S. capacity (1994) is ca 23,000 t yr.

Specifications for commercial acetonitrile are given in Table 4. The principal organic impurity in commercial acetonitrile is propionitrile; although small amounts of allyl alcohol also may be present.

Table 4. Specifications for Commercial Acetonitrile

Specification	Value
sp gr, at 20°C	0.783–0.787
distillation range, °C	
initial min	80.5
end pt, max	82.5
purity (min), wt %	99.0
acidity (max), wt %	0.05
copper (max), ppm	0.5
iron (max), ppm	0.5
water (max), wt %	0.3
color (max), Pt–Co	15

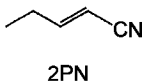
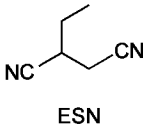
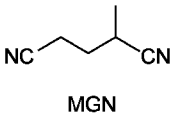
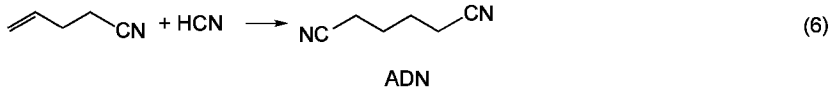
Shipping and Storage. The DOT/IMO classification for acetonitrile is class 3 hazard, UN No. 1648. It requires a FLAMMABLE LIQUID label on all containers and is in packing group II. For storage and piping at normal temperatures and pressures, aluminum or carbon steel may be used. Centrifugal pumps are preferred because the solvency of acetonitrile may affect the lubricant in positive-displacement pumps. All tanks, piping valves, and pumps should be electrically grounded. Fire protection equipment call be water spray, alcohol foam, CO₂, or dry chemical.

Health and Safety Factors. The following toxicities for acetonitrile have been reported: oral LD₅₀ (rats), 3030–6500 mg/kg; skin LD₅₀ (rabbits), 3884–7850 mg/kg; and inhalation LC₅₀ (rats), 7500–17,000 ppm (29). Humans can detect the odor of acetonitrile at 40 ppm. Exposure for 4 h at up to 80 ppm has not produced adverse effects. However, exposure for 4 h at 160 ppm results in reddening of the face and some temporary bronchial tightness.

Although acetonitrile has a low order of acute toxicity by ingestion, inhalation, and skin absorption, it can cause severe eye burns. In case of eye contact, eyes should be immediately flushed with water for at least 15 min and a physician should be consulted. In the event of a spill or leak, the spill should be contained, flooded with water, and disposed of according to local regulations. Acetonitrile is flammable (see Table 2) and must be kept away from excessive heat, sparks, and open flame. Associated fires can be extinguished using water spray, alcohol foam, CO₂, or dry chemical extinguishers. OSHA requires that an employee's exposure to acetonitrile in any 8-h shift does not exceed a time-weighted average of 40 ppm (70 mg/m³) in air (30).

Uses. Because of its good solvency and relatively low boiling point, acetonitrile is used widely as a recoverable reaction medium, particularly for the preparation of pharmaceuticals. Its largest use is for the separation of butadiene from C₄ hydrocarbons by extractive distillation (see AZEOTROPIC AND EXTRACTIVE DISTILLATION) (31). It also has been proposed for the separation of other olefins, eg, propylene, isoprene, allene, and methylacetylene from hydrocarbon streams (32–34). It is a superior solvent for polymers and can be used as a solvent for spinning fibers and for casting and molding plastics. It is used widely in spectrophotometry and electrochemistry. Since pure acetonitrile does not absorb uv light, it is commonly used as a solvent in high pressure liquid chromatography (hplc) for the detection of materials, eg, residual pesticides, in the ppb range; highly purified hplc grade acetonitrile is routinely available from suppliers.

Acetonitrile also is used as a catalyst and as an ingredient in transition-metal complex catalysts (35,36). There are many uses for it in the photographic industry and for the extraction and refining of copper and by-product ammonium sulfate (37–39). It also is used for dyeing textiles and in coating compositions (40,41). It is an effective stabilizer for chlorinated solvents, particularly in the presence of aluminum, and it has some application in



The first HCN addition (eq. 3) occurs at practical rates above 70°C under sufficient pressure to keep butadiene condensed in solution and produces the 1,4- and 1,2-addition products (3-pentenitrile [4635-87-4], 3PN, and 2-methyl-3-butenitrile [16529-56-9], 2M3BN) in a 2 to 1 ratio. Fortunately, thermodynamics favors 3PN (about 20:1) and 2M3BN may be isomerized to 3PN (eq. 4) in the presence of a nickel catalyst.

The selective addition of the second HCN to provide ADN requires the concurrent isomerization of 3PN to 4-pentenitrile [592-51-8], 4PN (eq. 5), and HCN addition to 4PN (eq. 6). A Lewis acid promoter is added to control selectivity and increase rate in these latter steps. Temperatures in the second addition are significantly lower and practical rates may be achieved above 20°C at atmospheric pressure. A key to the success of this homogeneous catalytic process is the ability to recover the nickel catalyst from product mixture by extraction with a hydrocarbon solvent. 2-Methylglutaronitrile [4553-62-2], MGN, ethylsuccinonitrile [17611-82-4], ESN, and 2-pentenitrile [25899-50-7], 2PN, are by-products of this process and are separated from adiponitrile by distillation.

Recent patent activity suggests that DuPont is developing a new generation of chelating diphosphite–nickel catalysts for this technology which are significantly more active than the monodentate phosphite based catalyst system used for the last two decades (61–64).

The Monsanto adiponitrile process, first commercialized in 1965 (65–67), involves the dimerization of acrylonitrile at the cathode in an electrolytic cell (eq. 7):



Small amounts of propionitrile and bis(cyanoethyl) ether are formed as by-products. The hydrogen ions are formed from water at the anode and pass to the cathode through a membrane. The catholyte that is continuously recirculated in the cell consists of a mixture of acrylonitrile, water, and a tetraalkylammonium salt; the anolyte is recirculated aqueous sulfuric acid. A quantity of catholyte is continuously removed for recovery of adiponitrile and unreacted acrylonitrile; the latter is fed back to the catholyte with fresh acrylonitrile. Oxygen that is produced at the anodes is vented and water is added to the circulating anolyte to replace the water that is lost through electrolysis. The operating temperature of the cell is ca 50–60°C. Current densities are 0.25–1.5 A cm² (see ELECTROCHEMICAL PROCESSING).

A typical composition of the catholyte is adiponitrile, 15 wt %; acrylonitrile, 15 wt %; quaternary ammonium salt, 39 wt %; water, 29 wt %; and by-products, 2 wt %. Such a solution is extracted with acrylonitrile and water, which separates the organics from the salt that can be returned to the cell. The acrylonitrile is distilled from the extract and the resultant residue consists of ca 91 wt % adiponitrile, which is purified further by distillation. The overall yield of acrylonitrile to adiponitrile is 92–95 %.

Production and Shipment. Estimated adiponitrile production capacities in the U.S. in 1992 were about 625 thousand metric tons and worldwide capacity was in excess of 10 metric tons. The DOT IMO classification for adiponitrile is class 6.1 hazard, UN No. 2205. It requires a POISON label on all containers and is in packing group III. Approved materials of construction for shipping, storage, and associated transportation equipment are carbon steel and type 316 stainless steel. Either centrifugal or positive displacement pumps may be used. Carbon dioxide or chemical-foam fire extinguishers should be used. There are no specifications for commercial adiponitrile. The typical composition is 99.5 wt % adiponitrile. Impurities that may be present depend on the method of manufacture, and thus, vary depending on the source.

Health and Safety Factors. See "General Health and Safety Factors." The following toxicities for adiponitrile have been reported: oral LD₅₀ (rats), 300 mg/kg; dermal LD₅₀ (rabbits), 2,134 mg/kg; and inhalation 4-h LC₅₀ (rats), 1.7 mg. NIOSH has proposed an exposure limit of 4 ppm as a TWA (68).

Uses. The principal use of adiponitrile is for hydrogenation to hexamethylene diamine leading to nylon-6,6. However, as a result of BASF's new adiponitrile-to-caprolactam process, a significant fraction of ADN produced may find its way into nylon-6 production. Adipoquanamine, which is prepared by the reaction of adiponitrile with dicyandiamide [461-58-5] (cyanoguanidine), may have uses in melamine–urea amino resins (qv) (see "Benzonitrile, Uses"). Its typical liquid nitrile properties suggest its use as an extractant for aromatic hydrocarbons.

α-Aminonitriles

α-Aminonitriles are compounds containing both cyano and amine substituents attached to the same carbon atom. They are versatile synthetic intermediates that are used to make aminoacids, agrichemicals, chelants, radical initiators, and water-treatment chemicals. In some cases, aminonitriles produced as intermediates are not isolated, but immediately further reacted, for example by hydrolysis, as is the case in producing ethylenediaminetetraacetate (EDTA) or nitrilotriacetate (NTA). Isolated and commercially available aminonitriles include 2-amino-2-methylpropanenitrile (aminoisobutyronitrile, AN-64) [19355-69-2], 2-amino-2-methylbutanenitrile (AN-67) [4475-95-0], 2-amino-2,4-dimethylpentanenitrile (AN-52) [26842-43-3], and 1-aminocyclohexane carbonitrile (AN-88) [5496-10-6]. The designation in parentheses arise from their identity as intermediates in the production of azo radical initiators (see below).

Historically these compounds have been made in two-step processes. For smaller volumes, reaction of an appropriate ketone or aldehyde with a cyanide salt followed by treatment with an ammonium salt proves satisfactory (Strecker synthesis). For larger volumes, treatment of the ketone or aldehyde with HCN to produce a cyanohydrin, followed by treatment with ammonia has been practiced. However, in 1990, DuPont began practicing a new one-step

process (69) in which the a ketone is treated simultaneously with both HCN and ammonia at 40–60°C. The new process, based on the understanding that a cyanohydrin is not on the reaction pathway (Fig. 1), is both faster and more selective. The uncatalyzed reactions are driven to completion by the presence of excess ammonia; water is a by-product. If excess ammonia is removed without separating the by-product water, the reaction may be reversed to produce the reagents. Therefore, these products are often stored and shipped in the presence of excess ammonia.

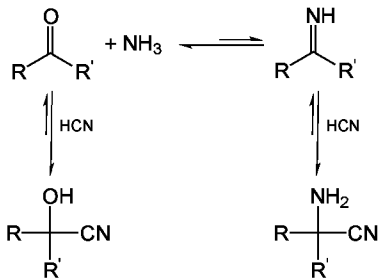


Fig. 1. Reactive pathway for α-aminonitriles synthesis.

Physical Properties. α-Aminonitriles are stable at modest temperatures (<70° C) in the absence of water; in the presence of water, they can degrade to their original constituents, ie, ketone (aldehyde), ammonia and hydrogen cyanide if insufficient ammonia is present. For this reason they are frequently stored in the presence of excess ammonia. Even in the presence of ammonia, aminonitriles begin to degrade at temperatures above 70°C, depending on the pressure of ammonia. The aminonitriles based on ketones described here are clear colorless liquids, but sometimes appear yellow to brown depending on the synthetic procedure and the amount of decomposition. They are soluble in polar organic solvents and in aromatic solvents. AN-64, AN-67, and AN-88 are soluble in water; AN-52 is insoluble in water. They have an ammonia-like odor. Vapor pressure of pure aminonitriles are AN-64, 4 kPa (30 torr) at 66°C; AN-67, 1.9 kPa (14 torr) at 68°C; AN-52, 66 Pa (0.5 torr) at 70°C and 267 Pa (2 torr) at 84°C; AN-88, 200–400 Pa (1.5–3 torr) at 82°C. Specific gravity is about 0.9 for AN-64, -67, and -52, and 1.03 for AN-88.

Shipment. The DOT IMO shipping information is shown in Table 6. Approved materials of construction for shipping, storage, and associated transportation equipment are lined carbon steel (DOT spec. 105 S 500W) and type 316 stainless steel. Water spray, carbon dioxide, chemical-foam, or dry-chemical fire extinguishers may be used.

Table 6. Aminonitrile Shipping Information

DOT IMO	AN-64 ^a	AN-67 ^a	AN-52 ^a	AN-88 ^a
hazard class	6.1	6.1	6.1	3
subs. hazard class	3	3	3	6.1
UN No.	2929	2929	2929	1992
label		POISON, FLAMMABLE LIQUID		
packing group	I	I	II	II

^a Shipping name: AN-64, Toxic Liquids, Flammable, Organic, N.O.S. (2-amino-2-methylpropanenitrile). AN-67, Toxic Liquids, Flammable, Organic, N.O.S. (2-amino-2-methylbutanenitrile). AN-52, Toxic Liquids, Flammable, Organic, N.O.S. (2-amino-2,4-dimethylpentanenitrile). AN-88, Flammable Liquids, Toxic, N.O.S. (1-amino cyclohexane carbonitrile).

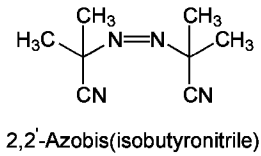
Health and Safety Factors. See "General Health and Safety Factors." As a class, alpha-aminonitriles are exceedingly acutely lethal from oral, inhalation, dermal, and ocular exposure. Some are known to cause osteolathyritic effects in experimental animals. The following toxicology data have been reported for the aminonitriles reported here. AN-64: oral LD₅₀ (rats) 45 mg kg, inhalation LC₅₀ (rats) 112 ppm; AN-67: inhalation LC₅₀ (rats) 111 ppm; AN-52: inhalation LC₅₀ (rats) 100 ppm, dermal LD₅₀ (rabbits) 90 mg kg; AN-88: oral LD₅₀ (rats) 200 mg kg, inhalation LC₅₀ (rats) 161 ppm, dermal LD₅₀ (rabbits) <200 mg kg. Additional data for these and other aminonitriles are available (12). These compounds are frequently stabilized with excess ammonia and thermal decomposition leads to the evolution of both ammonia and hydrogen cyanide. Accordingly, first-aid treatment should be consistent with both ammonia and cyanide exposure. In case of ingestion, drink two glasses of water and induce vomiting and call a physician (DO NOT give Syrup of Ipecac). Spills should be contained and treated with dry sodium bicarbonate (NaHCO₃) to absorb the spilled aminonitrile and make a dry solid at a ratio of three (3) pounds of NaHCO₃ per pound of aminonitrile. Transfer aminonitrile NaHCO₃ solids to plastic or metal drums for disposal. Firefighters may use water, carbon dioxide, dry-chemical or chemical foam extinguishers and should wear self-contained breathing apparatus.

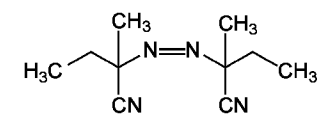
Uses. α-Aminonitriles may be hydrolyzed to aminoacids, such as is done in producing ethylenediaminetetracetate (EDTA) or nitrilotriacetate (NTA). In these cases, formaldehyde is utilized in place of a ketone in the synthesis. The principal use of the ketone-based aminonitriles described above is in the production of azobisnitrile radical initiators (see below). AN-64 is also used as an intermediate in the synthesis of the herbicide Bladex. Aminonitriles are also excellent intermediates for the synthesis of substituted hydantoins by reaction with carbon dioxide; however, this is not currently commercially practiced.

Azobisnitriles

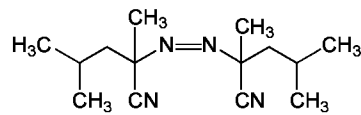
Azobisnitriles are efficient sources of free radicals for vinyl polymerizations and chain reactions, eg, chlorinations (see INITIATORS). These compounds decompose in a variety of solvents at nearly first-order rates to give free radicals with no evidence of induced chain decomposition. They can be used in bulk, solution, and suspension polymerizations, and because no oxygenated residues are produced, they are suitable for use in pigmented or dyed systems that may be susceptible to oxidative degradation.

DuPont (D) and Wako (W) produce several members of this class of compounds: 2,2'-azobis(isobutyronitrile) [78-67-1] (Vazo 64 (D); V-60 (W)); 2,2'-azobis(2-methylbutanenitrile) [13472-08-7] (Vazo 67 (D); V-59 (W)); 2,2'-azobis(2,4-dimethylpentanenitrile) [4419-11-8] [Vazo 52 (D); V-65 (W)]; 1,1'-azobis(cyanocyclohexane) [2094-98-6] [Vazo 88 (D)]; and 2,2'-azobis(4-methoxy-2,4-dimethylpentanenitrile) [15545-97-8] (V-70 (W)). They are crystalline solids that are produced by hypochlorite oxidation of α-aminonitriles (70). Physical properties are listed in Table 7 (71,72).

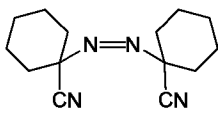




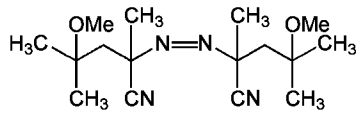
2,2'-Azobis(2-methylbutanenitrile)



2,2'-Azobis(2,4-dimethylpentanenitrile)



1,1'-Azobis(cyanocyclohexane)



2,2'-Azobis(4-methoxy-2,4-dimethylpentanenitrile)

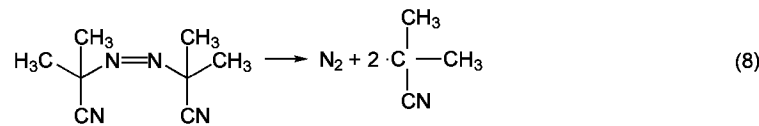
These compounds are essentially insoluble in water, sparingly soluble in aliphatic hydrocarbons, and soluble in functional compounds and aromatic hydrocarbons.

Table 7. Properties of Azobisnitriles

Property	V-70 ^b	Vazo 52 ^a	Vazo 64 ^a	Vazo 67 ^a	Vazo 88 ^a
		V-65 ^b	V-60 ^b	V-59 ^b	
CAS Registry Number	[15545-97-8]	[4419-11-8]	[78-67-1]	[13472-08-7]	[2094-98-6]
mol wt	308.4	248.4	164.2	192.3	244.3
mp, °C	50–96 d	45–70	100–103	48–52	113–115
specific gravity			1.128		
bulk density, kg m ⁻³ c		~450	~265	~450	~425
specific heat, kJ kgK ^d		~1.7	~1.7	~1.7	~1.7
heat of combustion, kJ kgmol ^d			5.05		
10-h half-life decomp temp, °C ^e	30	52	64	67	88
half-life (t _{1/2}), min					
log(t _{1/2}) = A(1/T) - B					
A		6767	7142	7492	7660
B		18.04	18.36	19.22	18.39
max storage temp, °C		10	24	24	35

^a Registered trademark of DuPont.
^b Wako name.
^c In form of white noodles.
^d To convert J to cal, divide by 4.184.
^e In toluene.
^f See text for structures and chemical names.

In solution, the azobisnitriles decompose on heating to form two free radicals with the liberation of nitrogen (eq. 8):



Because the decomposition is first order, the rate of free-radical formation can be controlled by regulating the temperature; equations relating half-life to temperature are provided in Table 7. These decomposition rates are essentially independent of the solvent (73).

Shipping and Storage. DOT/IMO shipping information for Vazo products are provided in Table 8. Vazo polymerization initiators must be stored out of the sun and away from heat in a cool, dry place. Since decomposition produces nitrogen and, consequentially, a pressure increase, these compounds should not be stored in glass or in any tightly closed containers other than the original shipping container. The maximum storage temperatures are provided in Table 7.

Table 8. Azobisnitrile Shipping Information

DOT/IMO	Vazo 64 ^a	Vazo 67 ^a	Vazo 52 ^a	Vazo 88 ^a
hazard class	4.1	4.1	4.1	4.1

UN No.	3234	3236	3236	3226
label		-----FLAMMABLE SOLID-----		
packing group	II	II	II	II

^a Shipping name (see text for structure):
Vazo-64, Self-Reactive Solid Type C, Temperature Controlled (2,2'-azodi(iso-butyronitrile)).
Vazo-67, Self-Reactive Solid Type D, Temperature Controlled (2,2'-azodi(2-methylbutyronitrile)).
Vazo-52, Self-Reactive Solid Type D, Temperature Controlled (2,2'-azodi(2,4-dimethylvaleronitrile)).
Vazo-88, Self-Reactive Solid Type D (1,1'-azodi(hexahydrobenzonitrile)).

Health and Safety Factors. Some of the Vazo products are mild skin or eye irritants in laboratory animals (Table 9) but none are skin sensitizers. In the absence of a polymerizable vinyl polymer, tetramethylsuccinonitrile [3333-52-6] (TMSN) is the principal decomposition product of Vazo 64. TMSN is highly toxic orally (rat oral LD₅₀ of 39 mg kg) and by inhalation (29). OSHA regulations require that an employee's exposure to TMSN in any 8-h shift does not exceed an 8-h time-weighted average of 0.5 ppm in air (3 mg m⁻³). Because both TMSN solid and vapor are capable of penetrating the skin and mucous membranes, control of vapor inhalation alone may not be sufficient to prevent absorption of an excessive dose.

Table 9. Toxicity of Vazo Compounds

	Skin		Mild eye Irritant	Inhalation, ^a mg m ⁻³	Oral ^b , mg kg
	Irritant	Sensitive			
Vazo-64	no	no	yes	950	450
Vazo-67	no	no	no	>8,900	982
Vazo-52	yes	no	yes		>5,000
Vazo-88	no		yes		>11,800

^a Vazo-64, ALC (approx lethal concentration); Vazo-67, LC₅₀.
^b Vazo-64, -52, -88, ALD (approx lethal dose); Vazo-67, LD₅₀.
^c See text for chemical names and structures.

All operations should be carried out with good ventilation and contact with eyes and skin should be avoided. In case of eye contact, the eyes should be flushed with water for 15 min and a physician should be consulted. Soap and water should be used to wash azobisnitriles from skin. If these compounds are inhaled, particularly the decomposition products of Vazo 64, the victim should be removed to fresh air and oxygen should be administered. If the victim is not breathing, cardiopulmonary resuscitation should be administered. A physician should be called in either case. Small quantities of waste Vazo should be disposed of by incineration, preferably by dissolving first in a waste liquid. Large quantities of waste Vazo should be returned to the manufacturer for disposal.

Uses. The azobisnitriles have been used for bulk, solution, emulsion, and suspension polymerization of all of the common vinyl monomers, including ethylene, styrene vinyl chloride, vinyl acetate, acrylonitrile, and methyl methacrylate. The polymerizations of unsaturated polyesters and copolymerizations of vinyl compounds also have been initiated by these compounds.

Benzonitrile

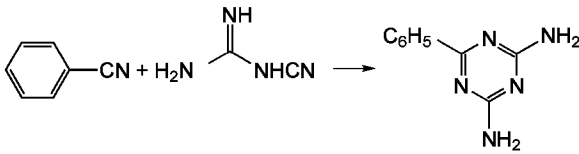
Benzonitrile [100-47-0], C₆H₅CN, is a colorless liquid with a characteristic almondlike odor. Its physical properties are listed in Table 10. It is miscible with acetone, benzene, chloroform, ethyl acetate, ethylene chloride, and other common organic solvents but is immiscible with water at ambient temperatures and soluble to ca 1 wt% at 100°C. It distills at atmospheric pressure without decomposition, but slowly discolors in the presence of light.

Table 10. Some Physical Properties of Benzonitrile

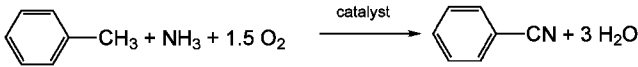
Property	Value
mol wt	103.12
bp (at 101.3 kPa = 1 atm), °C	191.1
freezing pt, °C	12.8
density (at 20°C), g cm ⁻³	1.0102
n _D ²⁰	1.5289
viscosity, mPas (cP)	1.24
heat of vaporization, mJ kg ^a	3.66
surface tension (at 20°C), mN m (dyn cm)	39.0
coefficient of expansion (per °C)	9 × 10 ⁻⁴
dielectric constant (at 25°C)	25.5
dipole moment (at 22°C), Cm ^b	13.144 × 10 ⁻³⁰
evaporation rate (butyl acetate = 100)	13
flash pt (COC) ^c , °C	79

^a To convert J to cal, divide by 4.184.
^b To convert Cm to D, divide by 3.336 × 10⁻³⁰.
^c COC = Cleveland open cup.

Like acetonitrile, benzonitrile is a powerful solvent for many inorganic and organic materials including some polymers. Inorganic salts that are soluble in benzonitrile at 25°C include aluminum chloride, ferric chloride, and silver nitrate. In chemical reactions, benzonitrile displays characteristics of both the nitrile group and the aromatic nucleus. It can be converted to a large number and variety of derivatives by simple syntheses; eg, by hydrolysis, it can be converted to either benzoic acid or benzamide. Hydrolysis in the presence of ZnCl₂ and HCl produces benzaldehyde, whereas alcoholysis leads to the formation of benzoic acid esters. Reduction produces benzyl- or dibenzylamine and reductive alkylation produces N-substituted benzamides. The aromatic ring can be halogenated or nitrated without the nitrile group being affected. Benzonitrile can be trimerized in the presence of a catalytic amount of base to 2,4,6-triphenyl-1,3,5-triazine (74). The most important reaction is with dicyandiamide to produce 2,4-diamino-6-phenyl-1,3,5-triazine (benzoguanamine) (75):

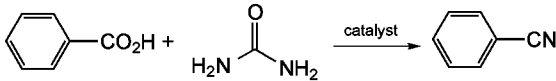


Benzonitrile can be produced in high yield by the vapor-phase catalytic ammoxidation of toluene (76):



An older route is based on just toluene and ammonia in the absence of air with separate catalyst regeneration (77).

In 1987, Toray Industries, Inc., announced the development of a new process for making aromatic nitriles which reportedly halved the production cost, reduced waste treatment requirements, and reduced production time by more than two-thirds, compared with the vapor-phase process used by most producers. The process involves the reaction of benzoic acid (or substituted benzoic acid) with urea at 220–240°C in the presence of a metallic catalyst (78).



A method for making benzonitrile by dehydrogenation of the Diels-Alder adduct of butadiene and acrylonitrile also has been described (79). Benzonitrile also can be made on a small scale by the dehydration of benzamide in an inert solvent with phosphorus oxychloride or benzenesulfonyl chloride and an organic amine (80,81).

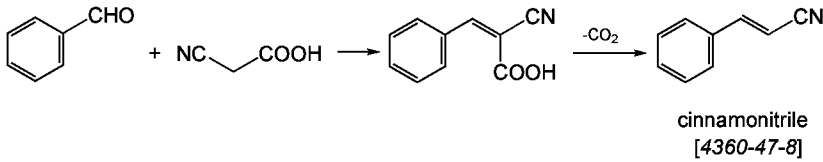
Shipping and Storage. The DOT hazard classification for BENZONITRILE is "Flammable", UN No. 2224. Carbon-steel drums and tanks may be used for storage.

Health and Safety Factors. See "General Health and Safety Factors." The following toxicities for benzonitrile have been reported: oral LD_{LO} (rats), 720 mg kg; skin LD₅₀ (rats), 1200 mg kg; and inhalation LC₅₀ (rats), 950 ppm 8 h.

Uses. The most important commercial use for benzonitrile is the synthesis of benzoguanamine, which is a derivative of melamine and is used in protective coatings and molding resins (see AMINO RESINS; CYANAMIDES). Other uses for benzonitrile are as an additive in nickel-plating baths, for separating naphthalene and alkylnaphthalenes from nonaromatics by azeotropic distillation (qv), as a jet-fuel additive, in cotton bleaching baths, as a drying additive for acrylic fibers, and in the removal of titanium tetrachloride and vanadium oxychloride from silicon tetrachloride.

Cyanoacetic Acid and Esters

The physical properties of cyanoacetic acid [372-09-8] and two of its ester derivatives are listed in Table 11 (82). The parent acid is a strong organic acid with a dissociation constant at 25°C of 3.36 × 10³. It is prepared by the reaction of chloroacetic acid with sodium cyanide. It is hygroscopic and highly soluble in alcohols and diethyl ether but insoluble in both aromatic and aliphatic hydrocarbons. It undergoes typical nitrile and acid reactions but the presence of the nitrile and the carboxylic acid on the same carbon cause the hydrogens on C-2 to be readily replaced. The resulting malonic acid derivative decarboxylates to a substituted acrylonitrile:



The methyl and ethyl esters of cyanoacetic acid are slightly soluble in water but are completely miscible in most common organic solvents including aromatic hydrocarbons. The esters, like the parent acid, are highly reactive, particularly in reactions involving the central carbon atom; however, the esters tend not to decarboxylate. They are prepared by esterification of cyanoacetic acid and are used principally as chemical intermediates.

Table 11. Some Physical Properties of Cyanoacetic Acid and Methyl and Ethyl Esters^a

	Cyanoacetic acid, NCCH ₂ COOH	Methyl ester, NCCH ₂ COOCH ₃	Ethyl ester, NCCH ₂ COOC ₂ H ₅
mol wt	85.06	99.09	113.11
bp, °C	108 ^b	206	208–210
mp, °C	67		22.5
n _D ²⁰		1.419	1.4177
flash pt, °C	107	110	110

^a Ref. 70.

^b At 2 kPa (15 mm Hg).

Health and Safety Factors. The following toxicities have been reported for cyanoacetic acid: oral LD₅₀ (rat) 1500 mg kg; subcutaneous LD_{LO} (rabbit), 1900 mg kg; and subcutaneous LD_{LO} (frog); 1300 mg kg (29). For ethyl cyanoacetate the following toxicities have been reported: interperitoneal LD₅₀ (mice), 750 mg kg; subcutaneous LD_{LO} (rabbits), 1500 mg kg; and subcutaneous LD_{LO} (frogs), 4000 mg kg.

Uses. Although cyanoacetic acid can be used in applications requiring strong organic acids, its principal use is in the preparation of malonic esters and other reagents used in the manufacture of pharmaceuticals, eg, barbitol, caffeine, and B vitamins (see ALKALOIDS; HYPNOTICS; VITAMINS). Cyanoacetic acid can be used for the preparation of heterocyclic ketones.

Isophthalonitrile

Isophthalonitrile [626-17-5] (1,3-dicyanobenzene, IPN), is a white solid which melts at 161°C and sublims at 265°C. It is slightly soluble in water but readily dissolves in dimethylformamide, N-methylpyrrolidinone and hot aromatic solvents. IPN undergoes the reactions expected of an aromatic nitrile, eg, hydrogenation of both nitrile groups and aromatic ring and chlorination to tetrachloroisophthalonitrile. IPN is prepared by vapor-phase ammoxidation of *meta*-xylene. The oral LD₅₀ for rats is 860 mg kg. It's principal use appears to be as an intermediate to amines. As a reagent, IPN can be used to convert

aromatic acids to nitriles in near quantitative yields (83).

2-Methylglutaronitrile

Methylglutaronitrile (2,3-dicyanobutane) *[4553-62-2]*, MGN, is a by-product of DuPont's adiponitrile process. The oral LD₅₀ (rats) is 400 mg kg (29). Some physical properties are listed in Table 12.

Table 12. Some Physical Properties of Methylglutaronitrile

Property	Values
mol wt	108.14
physical state	colorless liquid
bp, °C (at 101.3 kPa)	274
mp, °C	44
vapor pressure (kPa at 179°C)	8.1
specific gravity, at 25°C	0.95
flash pt, °C	98
flammable limits (in air), vol %	
lower	0.3
upper	3.25

Shipping and Storage. The DOT shipping name for MGN is "Toxic liquid, Organic, N. O. S. (Methylglutaronitrile)" and is in the hazard class 6.1, packing group III, UN 2810. It requires a "POISON" label. Carbon-steel drums and tanks may be used for storage.

Uses. Methylglutaronitrile is readily hydrogenated to give 2-methyl-1,5-pentanediamine (DYTEK A, MPMD), used as a comonomer in polyamide fibers and resins, as a curing agent for epoxy coatings, and as its isocyanates in specialty urethanes. A co-product of the DYTEK A process is 3-methylpiperidine which can be used to produce vulcanization accelerators for rubber curing or can be converted to 3-picoline, an intermediate used to make niacin and niacinamide. 3-Picoline may also be made by direct conversion of MGN or alternatively from DYTEK A.

Pentenitriles

Pentenitriles are produced as intermediates and by-products in DuPont's adiponitrile process. 3-Pentenitrile *[4635-87-4]*, is the principal product isolated from the isomerization of 2-methyl-3-butenitrile (see eq. 4). It is entirely used to make adiponitrile. *cis*-2-Pentenitrile *[25899-50-7]*, is a by-product isolated from the second hydrocyanation step. Some physical properties are listed in Table 13.

Table 13. Some Physical Properties of Pentenenitrile

Property	3-Pentenitrile	<i>cis</i> -2-Pentenitrile
mol wt	81.05	81.05
physical state	colorless liquid	colorless liquid
bp, °C	145	127
vapor pressure (kPa at 50°C)	3.3	6.3
specific gravity, at 20°C	0.83	
flash pt, °C	40	26

Shipping and Storage. The DOT shipping name for 2[3]-pentenenitrile is "Toxic liquid, Flammable, Organic, N. O. S. (2[3]-Pentenitrile)" and is in the hazard class 6.1, packing group II, UN 2929. It requires a "POISON, FLAMMABLE" label. Carbon-steel drums and tanks may be used for storage.

Health and Safety Factors. See "General Health and Safety Factors." The following toxicities for 2-pentenitrile have been reported: dermal, LD₅₀ (rabbit) >200 mg kg; oral LD_{LO} (rat) 450 mg kg; 4-hr Inhalation LC₅₀ (rats) 740 ppm (29).

Uses. 3-Pentenitrile, 3PN, is used entirely by the manufacturers to make adiponitrile. *cis*-2-Pentenitrile, 2PN, can be cyclized catalytically at high temperature to produce pyridine, a solvent and agricultural chemical intermediate. 2PN is also chlorinated to manufacture pentachloropyridine, an intermediate in the insecticide Dursban produced by Dow. Addition of ammonia to 2PN followed by reduction leads to 1,3-pentadiamine (Dytek ep), which is used as a curing agent for epoxy coatings and as a chain modifier in polyurethanes.

Fatty Acid Nitriles

Some of the physical properties of fatty acid nitriles are listed in Table 14 (see also CARBOXYLIC ACIDS). Fatty acid nitriles are produced as intermediates for a large variety of amines and amides. Estimated U.S. production capacity (1980) was >140,000 t yr. Fatty acid nitriles are produced from the corresponding acids by a catalytic reaction with ammonia in the liquid phase. They have little use other than as intermediates but could have some utility as surfactants (qv), rust inhibitors, and plasticizers (qv).

Table 14. Some Physical Properties of Fatty Acid Nitriles

Nitrile	CAS Registry Number	Bp, ^a °C	Freezing pt, °C	Iodine value
oleyl	<i>[112-91-4]</i>	>325	5	85 min
coco	<i>[61789-53-5]</i>			15 max
tallow	<i>[61790-28-1]</i>	330–360	4	44–60
stearyl	<i>[638-65-3]</i>	330–360	39	4 max

^a At 101.3 kPa.

DISCLAIMER

This article has been reviewed by the Office of Pollution Prevention and Toxics, U.S. Environmental Protection Agency, and approved for publication. Approval does not

signify that the contents necessarily reflect the views and policies of the Agency, nor does mention of commercial products or synthesis constitute endorsement or recommendation for use.

BIBLIOGRAPHY

"Nitriles" in *ECT* 3rd ed., Supplement Vol. pp. 888–909, by R. A. Smiley, E. I. du Pont de Nemours & Co., Inc.

1. D. H. R. Barton and W. D. Ollis, *Comprehensive Organic Chemistry*, Vol. 2, Pergamon Press, Oxford, U.K., 1979, pp. 528–562.
2. D. T. Moury, *Chem. Rev.* **42**, 192 (1948).
3. **U.S. Pat. 2,915,455** (Nov. 10, 1959), R. A. Smiley (to DuPont).
4. **U.S. Pat. 2,481,826** (Sept. 13, 1949), J. N. Cosby (to Allied Chemical).
5. M. C. Sze and A. P. Gelbein, *Hydrocarbon Process.* **55**, 103 (Feb. 1976).
6. **U.S. Pat. 3,496,215** (Feb. 17, 1970), W. C. Drinkard, Jr. and R. V. Lindsey, Jr. (to DuPont).
7. **U.S. Pat. 3,536,748** (Oct. 27, 1970), W. C. Drinkard, Jr. and R. V. Lindsey, Jr. (to DuPont).
8. C. Y. Wu and H. E. Swift, *179th American Chemical Society Meeting, Div. of Petroleum Chemistry Symposia Preprint*, Houston, Texas, Mar. 1980, p. 372.
9. C. A. Tolman, R. J. McKinney, W. C. Seidel, J. D. Druliner, and W. R. Stevens, *Adv. Catal.* **33**, 1 (1985).
10. C. De Bellefon and P. Flouilloux, *Catalysis Reviews*, **36**, 459–506 (1994).
11. *United States International Trade Commission Publication* 2933, Nov. 1995.
12. S. C. DeVito, in S. C. DeVito and R. L. Garrett, eds., *Designing Safer Chemicals: Green Chemistry for Pollution Prevention*, American Chemical Society Symposium Series 640, American Chemical Society, Washington, D.C., pp. 194–223.
13. **U.S. Pat. 3,362,889** (Jan. 9, 1968) J. F. Hannan (to Du Pont).
14. *Advances in Chemistry Series, Azetotropic Data, No. 6 and No. 35*, ACS, Washington, D.C., 1952 and 1962.
15. **U.S. Pat. 3,822,313** (July 2, 1974), J. R. Norell (to Phillips Petroleum Co.).
16. A. Buzas and co-workers, *Rev. Chim.* **22**, 656 (1971).
17. N. S. Kozlov and co-workers, *Dokl. Akad. Nauk BSSR* **2**(4), 326 (1977).
18. G. W. Gokel and co-workers, *Tetrahedron Lett.* **39**, 3495 (1976).
19. **U.S. Pat. 3,810,935** (May 14, 1974), W. Leimgruber and M. Weigele (to Hoffmann-LaRoche Inc.).
20. **U.S. Pat. 3,825,581** (July 23, 1974), W. A. Gay and D. F. Gavin (to Olin Corp.).
21. H. Felkin and G. Roussi, *C. R. Acad. Sci. Paris ser. C* **266**, 1552 (1968).
22. **U.S. Pat. 3,920,718** (Nov. 18, 1975), J. E. Nottke (to DuPont).
23. **U.S. Pat. 3,829,429** (Aug. 13, 1974), R. A. Clement (to DuPont).
24. **Fr. Pat. 2,121,106** (Sept. 22, 1972), (to Ajinomoto Co., Inc.).
25. T. Cairns and co-workers, *J. Am. Soc.* **74**, 5633 (1952).
26. I. A. Gunevich and co-workers, *Tr. Mosk. Khim. Technol. Inst.* **86**, 5 (1975).
27. A. A. Durgarryan and co-workers, *Arm. Khim. Zh.* **25**, 401 (1972).
28. **U.S. Pat. 3,281,450** (Oct. 24, 1966), P. J. Horvath (to B. F. Goodrich).
29. Registry of Toxic Effects of Chemical Substances (RTECS) On-Line Database. National Library of Medicine, Bethesda, Maryland.
30. *OSHA Standard 1910*, June 27, 1974.
31. H. D. Evans and D. H. Samo, *7th World Petrol. Congr. Proc.* **5**, 259 (1967).
32. **E. Ger. Pat. 91,480** (July 20, 1972), G. Hauthal and co-workers.
33. M. Enomotto and co-workers, *Kagaku Kogaku* **35**, 437 (1971).
34. **USSR Pat. 439,143** (Mar. 15, 1976), M. E. Aerov and co-workers.
35. A. Krause and T. Weiman, *Russkii Chem.* **40**, 1173 (1967).
36. **Ger. Offen. 2,237,704** (Feb. 15, 1973), L. W. Gosser and C. A. Tolman (to DuPont).
37. **Brit. Pat. 1,312,573** (Apr. 4, 1973), N. T. Notley.
38. D. M. Muir and A. J. Parker, *Adv. Entr. Metall. 3rd., Int. Symp.* 191 (1977).
39. **U.S. Pat. 3,607,136** (Sept. 21, 1971), R. A. Smiley and J. A. Vernon (to DuPont).
40. L. I. Primak and co-workers, *Tekst. Promst. (Moscow)* (2), 60–61 (1977).
41. **Fr. Pat. 1,471,321** (Mar. 3, 1967), (to Dunlop Co., Ltd.).
42. **Jpn. Kokai 76 04,107** (Jan. 14, 1976), T. Kita and M. Ishii.
43. W. S. Brud and co-workers, *6th Int. Cong. Essent. Oils*, 73 (1974).
44. *Chem. Week*, **158**, 10 (April 24, 1996).
45. **U.S. Pat. 5,495,016** (Feb. 27, 1996) G. Achhammer and E. Fuchs (to BASF).
46. **Ger. Pat. 848,498** (Sept. 4, 1952), K. Adam and co-workers (to BASF).
47. **Fr. Pat. 1,483,300** (June 2, 1967), K. Adam and co-workers (to BASF).
48. **Belg. Pat. 763,109** (Aug. 17, 1971) (to ICI).
49. **U.S. Pat. 2,284,525** (May 26, 1942), A. W. Larchar and co-workers (to DuPont).
50. **U.S. Pat. 3,696,153** (Oct. 3, 1972), B. J. Kershaw and co-workers (to DuPont).
51. **U.S. Pat. 3,821,305** (June 28, 1974) G. Bartalini and M. Gioggioli (to Montedison Fibre).
52. **Brit. Pat. 568,941** (Apr. 26, 1945), (to DuPont Co.).
53. **U.S. Pat. 3,299,116** (Jan. 17, 1967), R. Romani and M. Ferri (to Societa Rodiatoce).
54. **U.S. Pat. 3,153,084** (Oct. 13, 1964), T. M. Veazey and co-workers (to DuPont).
55. **Brit. Pat. 893,709** (Apr. 11, 1962), (to Chemstrand).
56. **U.S. Pat. 3,242,204** (Mar. 22, 1966), W. A. Lozier (to DuPont).
57. J. Szymanowski and A. Sobczynska, *Przem. Chem.* **66**, 373–377 (1987).
58. **U.S. Pat. 2,680,761** (June 8, 1954), R. H. Hallwell (to DuPont).
59. **U.S. Pat. 2,749,359** (June 5, 1956), W. H. Calkins and co-workers (to DuPont).
60. **U.S. Pat. 2,532,312** (Dec. 5, 1950), L. E. Romilly (to DuPont).
61. **U.S. Pats. 5,440,067** (August 8, 1995) and **5,449,807** (Sept. 12, 1995), J. D. Druliner (to DuPont).
62. **U.S. Pat. 5,512,695** (April 30, 1996) and **WO9528228** (Oct. 26, 1995), K. Kreutzer and W. Tam (to DuPont).
63. **U.S. Pat. 5,512,696** (April 30, 1996), K. Kreutzer and W. Tam (to DuPont).
64. **WO9514659** (June 1, 1995), K. Kreutzer, R. J. McKinney, and W. Tam (to DuPont).
65. M. M. Baizer, *J. Electrochem. Soc.* **111**, 215 (1964).
66. M. M. Baizer, *Chemtech*, 161 (1980).
67. D. E. Danly, *Chemtech*, 302 (1980).
68. R. Hartung, in G. D. Clayton and F. E. Clayton, eds., *Patty's Industrial Hygiene and Toxicology, Volume II, Part D*, John Wiley & Sons, New York, 1994,

pp. 3119–3172.

69. **U.S. Pat. 5,194,657** (March 16, 1993) R. J. McKinney and R. N. McGill (to DuPont).

70. **U.S. Pat. 2,711,405** (June 21, 1955), A. W. Anderson (to DuPont).

71. *VAZO*, Product information bulletin, DuPont.

72. *Azo Polymerization Initiators*, Product information Bulletin, Wako Pure Chemical Industries, LTD.

73. C. Walling, *J. Polym. Sci.* **14**, 214 (1954).

74. **U.S. Pat. 2,598,811** (June 3, 1952), J. E. Mahan and S. D. Turk (to Phillips Petroleum).

75. **Can. Pat. 545,630** (Sept. 3, 1957), R. T. Corkum and J. M. Salsburg (to American Cyanamid).

76. **German Offen. 2,314,151** (Oct. 18, 1973), R. K. Grasselli and J. L. Callahan (to Standard Oil, Ohio) (for example).

77. W. I. Denton and co-workers, *Ind. Eng. Chem.* **42**, 796 (1950).

78. *Comline Chem. Mater.* 21 (Feb. 9, 1987).

79. **Brit. Pat. 968,752** (Sept. 2, 1964), M. H. Richmond (to Distillers Co.).

80. **Can. Pat. 722,712** (Nov. 30, 1965) C. Herschmann (to Lonza Ltd.).

81. **Fr. Pat. 1,170,116** (Jan. 9, 1959), C. R. Stephens (to Charles Pfizer & Co.).

82. *Cyanoacetic Esters*, Technical Brochure, Kay-Fries, Inc. Montvale, N.J.

83. W. G. Toland and L. L. Ferstandig, *J. Org. Chem.* **23**, 1350 (1958).

Ronald J. McKinney
E. I. du Pont de Nemours & Co., Inc.

Stephen C. DeVito
U.S. Environmental Protection Agency

RECYCLING, PLASTICS

Of the 200 million tons of municipal solid waste collected in the United States in 1993 (1), 22° was recycled while 62° was placed in landfills and 16° incinerated (2). Plastics comprised 9.3° of these materials. The number of U.S. residential collection programs increased from 1,000 in 1988 to more than 7,000 involving more than 100 million people in 1993 (2). Approximate 1994 U.S. recycling rates are given in Table 1.

Table 1. U.S. Recycling Rates of Various Materials^a

Material	Recycling rate, °
plastics	1
paper and paper board	40
ferrous metals (iron and steel)	37
aluminum	30
glass	7
rubber and leather	3

^a Ref. 3.

Although the recycling rate for plastic appears low, U.S. plastics production in 1993 was 70 billion pounds, while 700 million pounds was recycled. Other estimates of the U.S. recycling rate have been as much as 3.5° (2,4). Plastics recycling is a worldwide phenomenon. Aided by mandatory recycling laws, Germany's plastics packaging recycling has increased from about 90 million pounds in 1992 to 1.1 billion pounds in 1995 (5). Japan has enacted its own Container and Packaging Law to promote plastics and paper recycling (6). Great Britain has enacted a mandatory target plastics recycling rate of 15° to be achieved in 2001 (7). Export of recovered plastics from the United States to Asia is a significant business for plastics brokers (8).

Both economic factors and governmental regulations are driving recycling (9). Energy costs associated with recycling are almost always less than in manufacture of products from virgin materials. Plastics recycling takes only 10–15° of the energy needed to refine petroleum and manufacture virgin resins. Incineration of plastics is a less efficient means of saving energy. For example, 100 lb (45.4 kg) of high density polyethylene has a fuel value of 20 × 10⁶ Btu (19 kJ). Recycling saves twice this, 40 × 10⁶ Btu (38 kJ). Life cycle analysis has been used to determine the most economically and environmentally acceptable method of using recovered plastics: mechanical recycling of the plastic, depolymerization to produce monomers, or incineration to produce energy (10). The most acceptable technology depends on the type of polymer and local business and environmental conditions.

Both economic and environmental factors have led to government regulations designed to promote recycling. In some areas, the number of landfill sites is becoming limited. Although the number of landfills in the United States is declining, the remaining sites are large, modern facilities. Concerns about landfill disposal costs are becoming less of a factor in promoting North American recycling. However, the effect of plastic wastes on the environment is a growing concern.

Separation of Commingled Materials

Random mixing of plastics leads to a significant adverse effect on properties. For example, mixing a few percent polypropylene in polyethylene leads to a significant reduction in tensile strength due to the formation of two immiscible phases having little adhesion (11). Hence different types of plastics must be separated from each other. Solid wastes, particularly from residential curbside collection programs, arrive at material recovery facilities (MRF) as a complex mixture. MRFs are typically built to process 100 to 500 tons of waste per day (1). Unit operations are summarized in Figure 1. The wastes are dumped on a tipping floor. There paper products are separated from metals and plastics. Metals and plastics, mostly containers, are pushed onto a conveyer belt. Two types of magnetic separators remove steel and aluminum from plastics and glass. Density differences or manual sorting are used to separate glass from plastics. The glass containers are hand sorted by color. The plastics are separated into individual polymer types of the MRF or in separate reclaiming facilities (1). Plastic bottles are classified into clear poly(ethylene terephthalate) (PET) soft drink bottles; green PET soft drink bottles; translucent high density polyethylene milk, water, and juice bottles; pigmented high density polyethylene detergent bottles, poly(vinyl chloride) water bottles, and food containers such as polypropylene ketchup bottles (1). Processing equipment capable of separating pigmented bottles from clear ones has been installed in some facilities (12).

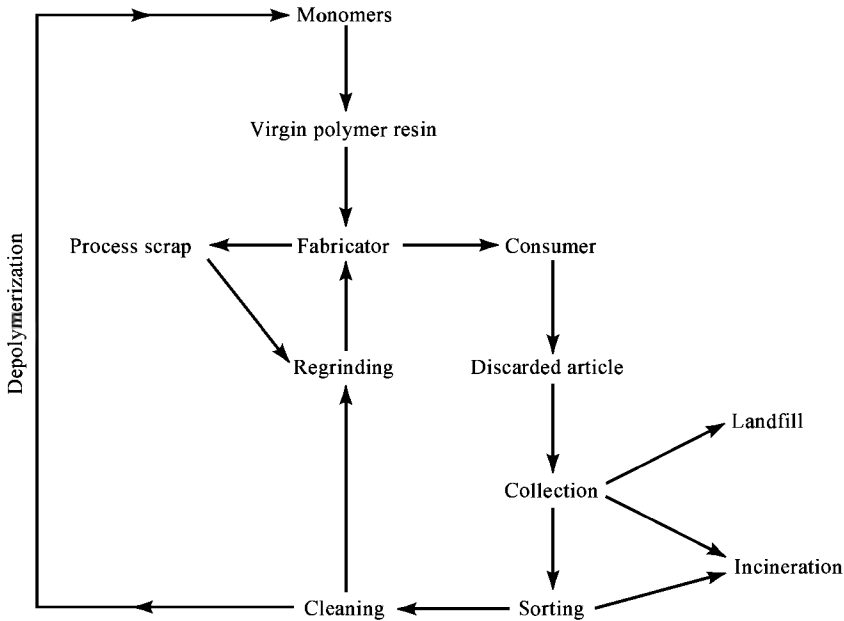


Fig. 1. Process steps in recycling.

When processing municipal solid wastes, an eddy current separation unit is often used to separate aluminum and other nonferrous metals from the waste stream. This is done after removal of the ferrous metals (see Fig. 1). The eddy current separator produces an electromagnetic field through which the waste passes. The nonferrous metals produce currents having a magnetic moment that is phased to repel the moment of the applied magnetic field. This repulsion causes the nonferrous metals to be thrown out of the process stream away from nonmetallic objects (13).

Another separation device that may be used is the mineral jig. This unit produces a loose vibrating bed of particles in a liquid medium. The vibrations segregate the solids into layers of density. The dense nonferrous metals, primarily lead, zinc, and copper are at the bottom while organics are at the top. The middle layer is primarily glass.

Separation of Impurities. After separation of the plastics, a number of impurities may still be present. This include inks used to print information and label onto plastics, other types of labels, wood, and dirt accumulated during use and disposal of the plastic. Washing technology has been used to remove inks, labels, and encrusted dirt from plastics, particularly bottles (14). A number of technologies have been used to separate other materials from plastics. Froth flotation has been used to separate poly(vinyl chloride) (PVC) from PET despite the similar densities of these polymers (15). In laboratory tests, froth flotation separated PVC, polycarbonate (PC), polyacetal, and poly(phenylene ether (PPE)) from each other. Wetting agents such as lignosulfonates, tannic acid, and saponin were required to promote the separation (16). Plastics may be separated from mixed bulk materials based on particle size (17). To separate poly(vinyl butyral) (PVB) in window glass from impurities, the polymer is melted and allowed to flow into supercritical carbon dioxide (18). Another technique to remove impurities from melted polymers is filtration (19).

Plastics

In May 1992, the U.S. Food and Drug Administration established the following guidelines to help assure the consumer safety of plastics recycling processes (20). Primary recycling is the recycling of plastics that are plant scrap and have not been sold for consumer use. Secondary recycling is the physical cleaning and processing of post-consumer plastic products. Tertiary recycling is the chemical treatment of polymers. This treatment is usually depolymerization to produce monomers which are purified and then polymerized to produce new polymer. Using tertiary recycling, materials such as fillers and fibers can be physically removed from the monomer. The monomers can also be purified by distillation and other processes prior to polymerization. The leading example of tertiary recycling is poly(ethylene terephthalate). Tertiary recycling also has been suggested for nylon from discarded carpets (21).

Residential collection programs indicate high collection rates for easily recognized types of containers (Table 2).

Table 2. Residential Recovery Rate by Package Type^a

Package type	Recovery, %
beverage bottles	65
liquid detergent bottles	50
other rigid containers	10
packaging film	5
average of all plastics	30

^a Ref. 1.

Sorted plastic packaging materials are shipped, usually in bales, to processing plants to be converted to polymer resins. The bales are broken and the bottles sorted to ensure that only one type of polymer is further processed. Processing consists of chopping and grinding the bottles into flakes. These flakes are washed. Processing steps such as flotation are used to remove polymeric contaminants from the flakes (15,16). The flakes are melted and converted into pellets.

For high value food packaging applications, minimal migration of contaminants into food products is critical. Currently the FDA requirement is a maximum 0.5 parts per billion (ppb) of noncarcinogenic compounds by dietary exposure (22).

Poly(ethylene terephthalate). About 1.6 billion pounds of PET are used in food packaging applications annually in the U.S. (23). About 42% of produced PET is recycled, mainly soft drink bottles (2). Cleaning of the recovered plastic comprises washing, rinsing, and drying. PVC is a common impurity in PET. Melting PET containing PVC will produce black spots due to charring of the PVC during processing to produce new bottles (24,25). Poly(vinyl chloride) (PVC) and PET have very similar density values; both will sink to the bottom of the water bath during rinsing. Therefore, it is difficult to separate the two polymers after the bottles have been ground into small particles (24). However, froth flotation has been shown to be an effective means of separating these two polymers (15).

For food applications, improved cleaning of PET produced by secondary recycling is needed. Supercritical fluid extraction using carbon dioxide (20) and solvents such as propylene glycol (26) have been proposed. High temperature and the use of vacuum to remove volatile impurities has also been suggested (27). Stripping of volatile components at temperatures above 160°C for three minutes has been reported (28). Application of multilayer approach, the manufacture of a bottle with an inner layer of recycled PET sandwiched between surface layers of virgin PET, is used commercially for soft drink applications (29).

Two PET tertiary recycling technologies are used commercially: methanolysis (30) and glycolysis (31). Both cleave the ester linkages in the polymer to form monomers. In methanolysis, methanol is used to cleave the polymer ester linkages producing stoichiometric amounts of dimethyl terephthalate and ethylene glycol (30,32). The ethylene glycol is separated and purified by distillation. The dimethyl terephthalate is purified by crystallization and distillation. Both bishydroxyethylterephthalate and oligomers are formed (31,32). These are recovered and purified by vacuum distillation and then polymerized in the presence of ethylene glycol to form PET.

Glycolysis is claimed to be somewhat less costly than methanolysis (33). Depolymerization is not taken completely to monomers (34). Rather, recovered PET is depolymerized to low molecular weight oligomers. Contaminants are removed using proprietary technology. The oligomers are then fed to a melt polymerization vessel in which PET is produced.

Hydrolysis yielding terephthalic acid and ethylene glycol is a third process (33). High temperatures and pressures are required for this currently noncommercial process. The purification of the terephthalic acid is costly and is the reason the hydrolysis process is no longer commercial.

Suitability for food contact is a critical property in selling recycled PET. The U.S. Food and Drug Administration (FDA) has issued letters of nonobjection to operators of methanolysis and glycolysis processes for recycling PET (36). Two-liter PET soda bottles containing recycled PET manufactured by glycolysis and methanolysis are widely available. The FDA has also issued a letter of nonobjection for a trilayer PET material having a middle layer sandwiched between two virgin layers of PET. Volatile materials associated with beverages and food can be removed efficiently from recovered PET by volatile stripping for three minutes at 160°C (35). However, if the air is recycled, volatile substances can be deposited as a film on the polymer particles.

Polyethylene. About 24% of produced HDPE is recycled in the U.S., mainly milk and water jugs and liquid laundry detergent bottles (2). Cleaning of the recovered plastic comprises washing, rinsing, and drying (24). Removing labels is the worst problem in washing and drying stages. Detergents are often used to improve the efficiency of label removal. Metal-foil labels can introduce metals into the polymer. When metal-foil labels are heat-sealed onto plastic, the only way to remove them is using an extrusion-melt filter. This leads to plugging of the filter screens causing more frequent changes and increasing production costs.

During the rinse cycle, polyethylene particles float to the surface of the water bath. The higher density PET and PVC particles sink to the bottom of

the bath and can be separated from the polyethylene.

HDPE has been molded into products such as plastic wood and the boards used in outdoor furniture. Other current uses include reuse container lids, truck bed liners, and pallets (36). Curbside residential recycling collection containers themselves often contain as much as 95% recycled HDPE (37). The U.S. Environmental Protection Agency has recommended that the U.S. government purchase shipping pallets containing 25–100% recycled polyethylene (38). This represents a sizable market as the U.S. Postal Service alone uses more than three million polyethylene pallets. Another large potential market in the U.S. is plastic and concrete railroad ties (39). About U.S.\$700 million is spent annually to replace 18–20 million of the 850 million wood and concrete ties currently in use. HDPE can also be shredded into small particles and blended with cement to create concrete (40). However, despite a low addition level (0–5% relative to cement), the HDPE had an adverse effect on cement compressive strength. An alkaline bleach treatment of the HDPE particles prior to blending reduced this adverse effect.

Blends of PET and HDPE have been suggested to exploit the availability of these clean recycled polymers. The blends could combine the inherent chemical resistance of HDPE with the processing characteristics of PET. Since the two polymers are mutually immiscible, about 5% compatibilizer must be added to the molten mixture (41). The properties of polymer blends containing 80–90% PET 20–10% HDPE have been reported (42). Use of 5–15% compatibilizer produces polymers more suitable for extrusion blow molding than pure PET.

Low density polyethylene has been pyrolyzed at 800°C to produce ethylene, propylene, and other light olefins in 75% yield (43).

Polypropylene. Polypropylene (PP) is used in packaging applications as films and in rigid containers. Battery cases could be considered another packaging application. Dead batteries are often collected at the point of sale of new batteries. In the U.S., some states have laws mandating this. Lead, acid, and plastics, particularly PP from battery casings is recovered and recycled (3). Care must be taken to limit worker exposure to lead during this process (44). PP is also recovered from bale wrap and other PP fabrics used for wrapping in the textile industry and from other containers (45).

Steps in polypropylene recycling include size reduction of grinding, washing, rinsing, and drying to remove contaminants and produce PP flakes (45). After extrusion, molten polymer is filtered through screen packs. The polymer may be separated into different melt flow ranges to produce more uniform product grades.

Polystyrene. Polystyrene (PS) is widely used in many packaging applications. These include injection molded products such as beverage containers, dairy product containers, and packaging for personal care products (46). Extruded solid sheet PS packaging products include salad boxes, dairy product containers, baked goods containers, and vending cups and lids. Extruded foam sheet PS packaging products include poultry and meat trays, produce trays, hinged lid containers, egg cartons, and foam cups. Blow and foam molded PS packaging products include vitamin bottles, loose fill packaging, and cushion packaging.

Polystyrene items separated from the solid waste stream are subjected to one or more of the following unit operations: densification (for foam), granulation to reduce particle size, washing, drying, extrusion, and pelletizing (46). The finished pellets have properties similar to the virgin resin. High density baling is used to increase the bulk density of polystyrene, often by a factor of two. Contaminants are more easily removed before this densification step than after. A demonstration plant chemically decomposes polystyrene to produce monomer (46). Polystyrene may be cracked at 130°C over sulfated zirconia to produce benzene (46).

Granulated polystyrene foam has been used as an additive in lightweight cement or as a soil additive to retain moisture and minimize compaction (46). Specially designed cup shredding machines for use with vending machines dispensing drinks in PS cups have been commercialized (7). However, recovery rates for other PS packaging products is significantly less than for easily recognized PS foam consumer product packaging (46).

Polystyrene has a high heating value, 46,000 kJ/kg compared to heating oil, 44,000 kJ/kg (46). Thus, incineration for its energy value is another possible application for recovered polystyrene.

Other Plastics. A relatively small amount of poly(vinyl chloride) goes into packaging applications and appears in municipal solid waste (25). The greatest concern with PVC is as a contaminant in other polymers being recycled, particularly PET. Approximately 12 million pounds of PVC was recycled in 1993, about half from packaging (25). Applications for recycled PVC include as an inner layer sandwiched between two virgin PVC layers in pipe and sheet for blister packaging and other packaging applications.

Polyurethane is pulverized to increase its bulk density, mixed with 30–80% of a thermoplastic molding material, gelled, and then granulated to give coated urethane foam particles 0.1 to 0.15 mm in size (48). The particle bulk density is three times that of the polyurethane, while the volume is 15% less. This material may be injection molded or extrusion molded into products (49). Other technologies for recycling polyurethanes have also been reported.

The recycling of engineering thermoplastics such as polyamides, ABS, and PTFE have been discussed (50). Property degradation as a result of use, recovery, and recycling is a concern.

Commingled Plastic Wastes. Owing to the property deteriorations that usually occur on polymer mixing (11), commingled plastics as useful and economic only for low value applications in which mechanical properties are not demanding. Such applications include park benches and parking barriers. Plastics in municipal solid waste streams are often contaminated with paper, which is difficult to separate from the plastic materials. If the cellulose fibers have a sufficient length, 0.2–2 mm (51), they can improve the mechanical properties of the plastic. Usually a reactive compatibilizer is required to improve the compatibility of the polymer phases and promote bonding of the cellulose to the plastics (52). One example cited is the addition of 30% cellulose fiber to a 70:30 mixture of LDPE and high impact polystyrene. An addition level of 30% maleic anhydride grafted styrene-ethylene-butylene-styrene block copolymer was used as the compatibilizer. This additive level was too high to be economic. However, the compatibilizer level was not optimized.

The supply of commingled plastics is much greater than the demand (53). Therefore, a critical issue in recycling commingled plastic wastes is the identification and separation of the plastics that are present. Near-infrared (900–1700 nm) spectroscopy has been proposed to identify polyethylene, poly(ethylene terephthalate), polypropylene, polystyrene, and poly(vinyl chloride) (PVC). A spectrograph with an InGaAs-array detector has been developed to record spectra from post-consumer packaging materials located on conveyer belts (54). Atomic absorption spectroscopy can be used when one of the polymers has a different atomic compositions than other polymers in a mixture (53). An example is the separation of poly(vinyl chloride) from polymers not containing chlorine.

The economics of recycling PET are more favorable than recycling HDPE. To increase the recycling of HDPE, the separation of bottles made of these two plastics could be omitted and a mixture processed. Coarse, light-colored powders of the two polymers have been prepared by an experimental solid state shear extrusion pulverization process (55). The powder has been successfully injection molded without pelletization.

The composition of nonmetal residues produced in shredding automobiles is summarized in Table 3.

Table 3. Representative Composition of Automobile Shredder Residue^a

Material	Wt, %
plastics	27
rubber	7
glass	16
textiles	12
fluids	17
other	21

^a Ref. 3.

Each vehicle generates 500–800 pounds of residue. The annual U.S. total is about 3.5 million tons or about 1.3% of the municipal solid waste generated annually (3). The mixture is too complex to separate and recycle. Depending on the amount of glass, water, metal, and dirt present, the residue

has a heating value of 4,800–6,800 Btus per pound (3,56). Incineration reduces residue weight by 50° and volume by 80° (4).

The thermal degradation of mixtures of the common automotive plastics polypropylene, ABS, PVC, and polyurethane can produce low molecular weight chemicals (57). Composition of the blend affected reaction rates. Sequential thermolysis and gasification of commingled plastics found in other waste streams to produce a syngas containing primarily carbon monoxide and hydrogen has been reported (58).

One alternative to identifying and separating different types of plastics is using commingled plastics directly. Since the composition and physical properties of commingled plastics can vary from day to day, applications are limited. One such product is a building material containing Portland cement as the binder, a filler (sand, gravel, or stone), and a plastic with a maximum particle size of 5–10 mm (59). The ratio of binder to aggregate is 1:4–8 while the volume ratio of binder + filler to plastic is 3–9:1. Concrete made using this material performs as well as standard concrete not containing plastic. In Germany, mixed plastics have been blended with pulverized fuel ash and the mixed used to manufacture fencing and posts (60).

Styrene block copolymers have been used as compatibilizers for mixed plastics to permit their processing for applications such as those outlined earlier (52,61).

Laboratory tests indicated that gamma radiation treatment and cross-linking using triallylcyanurate or acetylene produced a flexible recycled plastic from mixtures of polyethylene, polypropylene, general-purpose polystyrene, and high impact grade PS (62).

Another alternative to separating commingled plastics is advanced waste recycling. This is the high temperature–high pressure conversion of plastic wastes to form petrochemical process streams. Research is in progress to determine the conditions that will favor conversion of commingled plastic wastes to certain types of chemical feedstocks including synthesis gas (hydrogen + carbon monoxide), hydrogen, crude pyrolysis oil (containing benzene, toluene, and xylene), olefins, and oxygenates such as methanol, esters, and methyl formate (63). This technology has also been evaluated for producing fuels: medium BTU gas for boilers, and liquid fuels such as diesel oil. None of these processes is currently economic.

Thermal cracking of commingled plastics can produce an excellent feed for steam crackers and catalytic crackers. During steam cracking, feed from commingled plastics produced higher yields of ethylene (34° vs 28°), propylene (17° vs 15°), and butylene (12° vs 7°) than did the usual naphtha feed (64). During catalytic cracking, feed generated from plastics provided an 86° yield of naphtha grade product compared to a 62° yield from vacuum oil.

The Conrad recycling process utilizes an auger kiln to apply heat to plastics in the absence of oxygen (65). Feed preparation using the Conrad process requires minimal plastic particle size reduction, washing, and removal of nonplastic contaminants (66). Granulated plastics are introduced into a retort in the absence of oxygen using a rotary air lock. If the plastic particles are in their original form, they are introduced into the retort using a ram feeder. The plastics melt after entering the hot retort. An auger keeps the molten mass moving. Thermally promoted carbon–carbon bond cleavage occurs. As depolymerization proceeds, volatile products are produced and swept out of the reactor. Interestingly, mixing polymers seems to improve thermal cracking of results. A 1:1 mixture of polypropylene and polyethylene cracked at a lower temperature than did polyethylene alone and provided a narrower mixture of products than either polymer did separately (67).

Results obtained for two mixed plastics are summarized in Table 4. A balance exists between process temperature, plastics feed rate, and product yields (67). For example, lower temperatures increase wax formation due to incomplete depolymerization. Slower feed rates and increased residence times reduce wax formation and increase the yield of liquids. The data summarized in Table 4 illustrate that the addition of PET to a HDPE:PP:PS mixture changes the performance of the Conrad process. Compared to the reference HDPE:PP:PS mixture, increased amounts of solids are formed. These are 95° terephthalic acid and 5° mono- and bis-hydroxyethyl esters. At higher temperatures, apparently enough water remains to promote decarboxylation. In contrast, the addition of LDPE or PS to the mixture had little effect on its behavior in the Conrad process.

Table 4. Gas Chromatographic Analysis of the Results of Cracking Polymer Mixtures Using a Conrad Unit^{a,b}

Analysis	60:20:20 HDPE:PP:PS	20:48:16:16 PET:HDPE:PP:PS
liquid yield, wt °	73	55
gas yield, wt °	27	38
solids yield, wt °		7
	<i>Partial oil analysis, wt °</i>	
aliphatics, carbon number <10	9.4	21.3
11–15	16.6	17.6
16–20	9.9	9.9
21–25	3.5	4.7
26–30	1.7	2.2
31–40 +	1.5	2.2
aromatics		
benzene	2.3	1.4
toluene	11.0	3.8
ethylbenzene	5.7	2.6
xylene		1.1
other alkyl benzenes		8.4
styrene	17.0	8.2
naphthalenes		3.7
unidentified		12.3

^a Ref. 67.

^b Oven temperature = 649° C; Auger temperature = 527° C for the HDPE:PP:PS mixture, 479° C for the PET:HDPE:PP:PS mixture.

Fiber-Reinforced Plastics and Composites. It is usually too expensive to separate fillers and fibers from recovered polymers. Hence, the recycled use of these polymers must tolerate the presence of fillers or fibers (53). Thermoset matrix composites are ground and used as filler for polymers. Remolding is usually by injection or compression molding (68). Fiber-reinforced plastics are recycled primarily from old automobiles and electrical equipment (casings and various plastic parts). Glass fiber-reinforced plastics have been made into sheet-molding compounds and bulk-molding compounds (69). However, the economical recycling of fiber-reinforced plastics remains a challenge. Dynamic mechanical thermal analysis is said to determine heat resistance, impact resistance, and stiffness of glass-reinforced plastic before and after recycling (70). Thus, it could serve as a tool to determine the suitability of a glass-reinforced plastic for recycling.

Economics and Statistics

Costs of various waste disposal methods are summarized in Table 5.

Table 5. Estimated U.S. Processing Costs of Waste Disposal Method^a

Method	Cost estimate, \$ /t
--------	----------------------

landfilling	30 ^b
incineration for energy recovery	100
recycling of PET and HDPE bottles	100–150

^a Ref. 71.
^b Cost varies with the location of the landfill.

Polymer recycling process costs for various operations are summarized in Table 6.

Table 6. Approximate Polymer Recycling Costs^a

Process step	Costs	
	\$ per pound	%
collection	0.10	27
sorting	0.12	32
<i>subtotal cost</i>	<i>0.22</i>	<i>59</i>
grinding–cleaning	0.15	41
<i>Total</i>	<i>0.37</i>	<i>100</i>

^a Ref. 1.

Of course, the benefits gained by recycling are also important. These are summarized in Table 7 for PET and HDPE.

Table 7. Economic Benefits of Recycling^a

Polymer	Value per ton, \$	
	Incineration and energy recovery ^b	Recycling
PET	28–140	470
HDPE	312	120

^a Based on \$0.14 per kW h of energy.
^b

U.S. plastics production in 1993 was about 70 billion pounds (1). Plastics recycling continues to increase from its current value of 3.5% in the U.S. (2,4). However, recycling rates of some packaging products are much higher than this. The recycling rate of PET from soft drink bottles was 42% in 1993 while that of HDPE from milk and water jugs was 24% (2). This translates to 448 million pounds of PET and 450 million pounds of HDPE (1). For example, the process cost of recycling PET and HDPE bottles has been given as U.S.\$100–150 per ton (71). The value of the PET and HDPE produced from recycled materials is U.S.\$470 and U.S.\$120 per ton, respectively. These recycling processes can be profitable. The favorable economics of PET recycling have been attributed in part to forward integrated PET recyclers consuming their own product to make bottle resin (24).

Process costs for the methanolysis and glycolysis of PET to produce monomers are similar. Cost estimates are summarized in Table 8.

Table 8. Estimated Costs of PET Methanolysis Based on 1993–1994 Data^a

Item	Cost per pound of PET, \$
feedstock	0.30
conversion and handling	0.20
capital costs	0.15

^a Ref. 72.

By comparison, the 1993–1994 cost (per pound of PET) of dimethyl terephthalate and ethylene glycol made from petroleum was \$0.35.

However, recycling of many other plastics remains uneconomic (1). This is reflected in a number of companies closing plastics recycling operations in the mid-1990s (73). The costs of recycling commingled plastics has been estimated at U.S.\$1700 per ton (72). This is ten times more expensive than recycling easily separated homogeneous products such as PET and HDPE. In Germany, federal mandates require recycling of 60% of all plastic packaging. German projects to recycle over one billion pounds a year of plastics have been announced (74,75). These processes will use high temperature–high pressure processes to depolymerize commingled plastics to produce petrochemical feedstocks. These processes are not economic (65,76).

Few of the products in which polypropylene are used can be recovered in commercial qualities. An exception is battery casings. In 1994, U.S. capacity for recycling polypropylene from battery casings was 265 million pounds annually (45). About 75% of the recovered polypropylene is used in new battery cases which have a recycled PP content of about 50%. Total polypropylene recycling capacity was 350 million pounds annually for an operating rate estimated at 71–90% (45).

In 1993, over 41 million pounds of polystyrene was recycled into new plastics products in the U.S. (77). For commingled plastics, gasification comes closest to competing with low cost landfilling (57).

Price swings, particularly in the PET and HDPE markets have contributed to a retrenchment in the U.S. plastics recycling industry in 1995–1997 (78). In 1994 in the U.S., recycled PET, HDPE, LDPE, and PS had a 16–46% cost advantage (4). This cost advantage largely disappeared by 1996. Bureau of Labor Statistics data indicate U.S. plastics prices in mid-1997 are seven percent below those of mid-1995 after being more than ten percent less in 1996 (79). These lower prices make it more difficult for recycled plastics to compete with virgin resins in the absence of legislation mandating plastics recycling.

BIBLIOGRAPHY

"Recycling, Plastics" in *ECT* 3rd ed., Vol. 19, pp. 993–1002, by H. Alter, Chamber of Commerce of the United States.

- R. G. Saba and W. E. Pearson, in C. P. Rader, S. D. Baldwin, D. P. Cornell, G. B. Sadler, and R. F. Stockel, eds., *Plastics, Rubber and Paper Recycling: A Pragmatic Approach*, American Chemical Society, Washington, D.C., 1995, Chapt. 2, pp. 11–26.
- Franklin Associates Limited, "Characterization of Municipal Solid Waste in the United States, 1994 Update," Report No. EPA 530-94-042, Nov. 1994.

3. R. A. Pett, A. Golovny, and S. S. Labana, in Ref. 1, pp. 47–61.
4. C. P. Rader and R. F. Stockel, in Ref. 1, Chapt. 1, p. 3.
5. D. Castle, *Packag. Week* **12**(19), 3 (Oct. 10, 1996).
6. *Jpn. Chem. Week* **37** (1898), 1–2 (Oct. 24, 1996).
7. *Plast. Rubb. Wkly* **1659**, 8 (Oct. 25, 1996).
8. R. King, *Waste News* **1**(35), 14 (Apr. 29, 1996).
9. S. K. Adams and J. C. Even, in R. E. Landreth and P. E. Rebers, eds., *Municipal Solid Wastes* Lewis, Boca Raton, Fla., 1997, pp. 11–53.
10. K. H. Feuerherd, *Kunststoffe Plast. Europe* **86**(2), 19–20 (Feb. 1996). Translated from *Kunststoffe* **86**(2), 198–201 (1996).
11. R. S. Stein, in G. D. Andrews and P. M. Subramian, eds., *Emerging Technologies in Plastics Recycling*, American Chemical Society, Washington, D.C., ACS Symposium Series, 1992, Vol. 513, pp. 40–48.
12. B. Shearer, *Chem. Mark. Reporter* **249**(22), SR9 (May 27, 1996).
13. E. Schloeman and D. B. Spencer, *Resource Recovery Conserv.* **1**, 151 (1975).
14. W. Michaeli and M. Bittner, in J. Brandrup, ed., *Recycling and Recovery of Plastics* Hanser Publications, Munich, 1996, pp. 236–241.
15. R. Buchan and B. Yarar, *Min Eng.* **48**(11), 69–72 (1996).
16. J. Shibata, S. Matsumoto, H. Yamamoto, E. Kusaka, and Pradip, *Int. J. Miner. Process* **48**(3–4), 127–134 (1996).
17. W. Michaeli and M. Bittner, in Ref. 14, pp. 231–235.
18. PCT International Application WO 9705194 (Feb. 13, 1997) (to Cf Technologies, Inc.).
19. F. Hensen, in Ref. 14, pp. 291–296.
20. **U.S. Pat. 4,764,323** (Aug. 16, 1988), A. Ghatta (to Cobarr S.p.A.).
21. A. L. Bisio and M. Xanthos, eds., *How to Manage Plastic Waste: Technology and Market Opportunities*, Hanser Publishers, New York, 1995.
22. *Fed. Reg.* **58**, 52719–52729 (Oct. 12, 1993).
23. *Mod. Plast.* **71**, 73–94 (Jan. 1994).
24. W. K. Atkins, in Ref. 1, pp. 104–112.
25. R. A. Burnett, in Ref. 1, pp. 97–104.
26. **U.S. Pat. 4,680,060** (July 14, 1987), A. S. Gupta and J. T. Camp (to The Coca-Cola Company).
27. *Points to Consider for the Use of Recycled Plastics in Food Packaging: Chemistry Considerations*, U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, Publication HFS-245, Washington, D.C., April 1992.
28. D. E. Pierce, D. B. King, and G. D. Sadler, in Ref. 1, pp. 458–471.
29. T. H. Begley and H. C. Hollifield, *Food Technol.* **47**, 109–112 (November 1993).
30. D. Gisstis, *Makromol. Chem. Makromol. Symp.* **57**, 185 (1992).
31. G. Bauer in K. J. Thome-Kozmenschky, ed., *Recycling of Wastes*, EF-Press, Berlin, 1989, p. 293.
32. F. L. Bayer, D. V. Myers, and M. J. Gage, in Ref. 1, pp. 152–160.
33. D. E. Pierce, D. B. King, and G. D. Sadler, in Ref. 1, pp. 459–471.
34. E. N. Nowak and R. M. Oblath, in Ref. 1, pp. 404–417.
35. G. Sadler, in Ref. 1, pp. 380–388.
36. P. Post, *Waste News* **2**(19), 19 (July 15, 1996).
37. *Plast. Rub. Wkly.* 1639, 11 (June 7, 1996).
38. R. King, *Plast. News* **8**(27), 1–4 (Sept. 2, 1996).
39. R. King, *Plast. News* **8**(37), 50 (Nov. 11, 1996).
40. T. R. Naik, S. S. Singh, C. O. Huber, and B. S. Brodersen, *Cem. Concr. Res.* **26**(10), 1489–1492 (1996).
41. S. Jabarin and P. Sambaru, *Polym. Eng. Sci.* **33**(13), 827 (mid-July 1993).
42. S. A. Jabarin and V. V. Bhakkad, in Ref. 1, pp. 113–138.
43. S. F. Sodero, F. Bernuti, and L. A. Behie, *Chem. Eng. Sci.* **51**(11), 2805–2810 (1996).
44. J. M. Dave and A. M. Khan, in B. L. Johnson, C. Xintaris, and J. S. Andrews, eds., *Proceedings of the 2nd International Congress on Hazardous Wastes: Impacts on Human Ecology and Health* Princeton Scientific, Princeton, N.J., 1997, pp. 215–222.
45. R. M. Prioleau, in Ref. 1, pp. 80–88.
46. D. A. Thomson, in Ref. 1, pp. 89–96.
47. R. Lin and R. L. White, *Prepr. Pap.-Am. Chem. Soc. Div. Fuel Chem.* **41**(4), 1165–1169 (1996).
48. **Eur. Pat. Appl. EP 749,818** (Dec. 27, 1996), S. Nishibori and T. Kajiware (to Ein Engineering Co., Ltd.).
49. E. Weigand, in Ref. 14, pp. 702–716.
50. S. Akhtar and V. Kumar, *Pop. Plast. Packag.* **41**(3), 64–66 (1996).
51. A. J. Mitchell, J. E. Vaughan, and D. Willis, *Appl. Polym. Sci.* **22**, 2047 (1978).
52. P. Gatenholm, P. Hedenberg, and C. Klason, in Ref. 1, pp. 367–377.
53. R. S. Stein, in Ref. 1, pp. 27–46.
54. R. Feldhoff, D. Wienke, K. Cammann, and H. Fuchs, *Appl. Spectrosc.* **51**(3), 362 (1997).
55. K. Khait, *J. Vinyl Addit. Technol.* **2**(4), 345–348 (1996).
56. W. S. Hubble, I. G. Most, and M. R. Wolman, *Investigation of the Energy Value of Automobile Shredder Residue*, Publication DOE ID 12551-1, U.S. Department of Energy, Washington, D.C., Aug. 1987.
57. M. Day, J. D. Cooney, C. Klein, J. Fox, in *Polymer Durability*, Advances in Chemistry Series No. 249, American Chemical Society, Washington, D.C. 1996, pp. 47–57.
58. G. Mackey, in Ref. 1, pp. 160–169.
59. **Can. Pat. Appl. 2,126,224** (Oct. 4, 1996), J. P. Sawyers (to John P. Sawyers).
60. W. W. Todd, *Waste Management* **16**(1–3), 203–207 (1996).
61. D. Groote, P. Godard, and G. Vanhaeren, *Chim. Nouv.* **14**(56), 1635–1642 (1996).
62. Y. Fujii and M. Nomura, *Bull. Res. Lab. Nucl. React.* **20**, 72–74 (1996).
63. S. J. Pearson, G. D. Kryder, R. R. Kopang, and W. R. Seeker, in Ref. 1, pp. 183–193.
64. *Eur. Chem. News*, **62** 27 (Feb. 15, 1993).
65. **U.S. Patent 4,412,889** (1982), R. C. Oeck.
66. M. W. Meszaros, in Ref. 1, pp. 172–182.
67. *Thermal Recycling of Plastics, Final Report to the American Plastics Council*, Energy & Environmental Research Center, Grand Forks, N.D., March 1994.
68. J. Powell, *Resource Recycling*, 52 (May 1993).
69. J. M. Henshaw, W. Han, and A. D. Owens, *J. Thermoplast. Composite Mat.*, **9**(1), 4–20 (Jan. 1996).
70. P. Schaefer in Ref. 14, pp. 702–716.
71. J. H. Song, *J. Vinyl Addit. Technol.* **2**(3), 221–223 (1996).
72. J. W. Porter, *Recycling at the Crossroads*, Porter & Associates, Sterling, Va., 1993.
73. J. R. Ellis, in Ref. 1, pp. 62–69.

74. E. M. Kirschner, *Chem. Eng. News* **74**(45), 19–22 (Nov. 4, 1996).
75. P. Mapleston, *Modern Plastics* **70**, 53 (Nov. 1993).
76. *Eur. Chem. News*, **63** 37 (April 25, 1994)
77. E. Johnson, *Eur. Chem. News*, **65**(1700), 18 (Jan. 22–28, 1996).
78. Recycling Polystyrene, *Polystyrene Packaging Council* **5**(1), 2 (1994).
79. D. Smock, *Plasit. Formulating Compound*. **2**(1), 10–11 (Jan.–Feb. 1996).
80. D. M. Kiefer, *Today's Chemist at Work* **6**(5), 35–36 (May 1997).

John K. Borchardt
Shell Chemical Company

SEMICONDUCTORS, ORGANIC

Semiconductors are materials that are characterized by resistivities intermediate between those of metals and of insulators. The study of organic semiconductors has grown from research on conductivity mechanisms and structure–property relationships in solids to include applications-based research on working semiconductor junction devices. Organic materials are now used in transistors, photochromic devices, and commercially viable light-emitting diodes, and the utility of organic semiconductors continues to increase.

The study of organic semiconductors and conductors is highly interdisciplinary, involving the fields of chemistry, solid-state physics, engineering, and biology. This article provides a treatment of the theoretical aspects of organic semiconductors as well as an overview of recent advances in the field and the uses of these materials based on their conductive and optical properties.

Theory

The theory of conduction in organic semiconductors conveniently begins with a discussion of bonding in extended solids, since the nature of conduction is intimately related to the extent of delocalization of orbitals in the material to be studied. For inorganic conducting solids, the orbital overlap of the atoms in the crystal lattice results in the atomic energy levels spreading out to form energy bands rather than discrete levels. These energy bands allow for motion of charge carriers over many lattice sites without interruption by trapping events. The interactions between molecules in a molecular organic solid similarly leads to the formation of energy bands, with the extent of interaction between molecules (the extent of overlap of interacting orbitals) determining the width of the bands.

The distinctions between metals, semiconductors, and insulators are based on the band structure of the materials as well as on the electron occupancy of these bands. Figure 1 is a diagram of energy bands and occupancies for various classes of solids. As can be seen in Figure 1, insulators have a filled band formed from the valence orbitals (a "valence band") with a higher lying unfilled band formed from higher energy orbitals (the "conduction" band). The region between these has no allowed states, and the energy difference required to promote an electron from the valence to the conduction band is termed the band gap of the material. For an insulator, the band gap is large, typically >4 eV. Since all of the levels of the valence band are filled, no electrons are able to carry current.

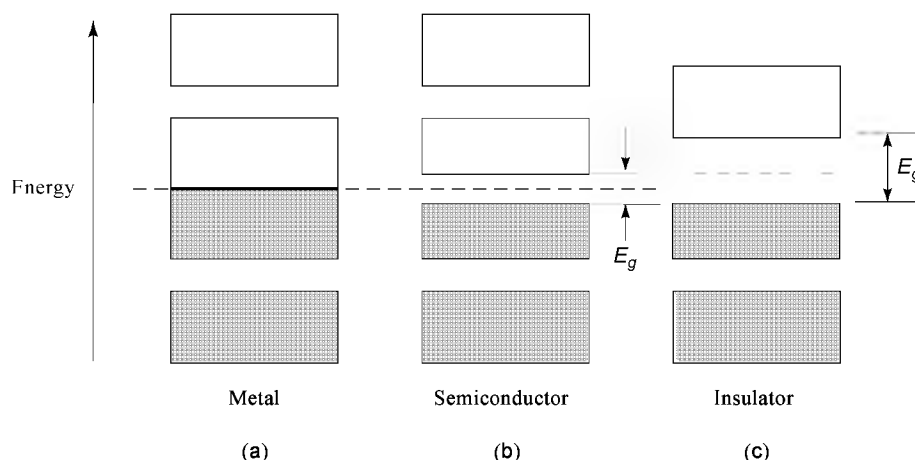


Fig. 1. Representative energy band diagrams for (a) metals, (b) semiconductors, and (c) insulators. The dashed line represents the Fermi Level, and the shaded areas represent filled states of the bands. E_g denotes the band gap of the material.

Metallic behavior is observed for those solids that have partially filled bands (Fig. 1b), that is, for materials that have their Fermi level within a band. Since the energy bands are delocalized throughout the crystal, electrons in partially filled bands are free to move in the presence of an electric field, and large conductivity results. Conduction in metals shows a decrease in conductivity at higher temperatures, since scattering mechanisms (lattice phonons, etc) are frozen out at lower temperatures, but become more important as the temperature is raised.

Semiconductors show thermally activated conductivity. The reason can be seen from Figure 1c. The band gap of a semiconductor is small relative to that of an insulator. Similar to insulators, pure intrinsic semiconductors have a filled valence band and unfilled conduction band. However, this is strictly true only at absolute zero. At higher temperatures, the energy of the surroundings kT can become large enough to thermally excite electrons from the valence to the conduction band, with the result that semiconductors have partially unfilled orbitals, giving rise to conduction. Typical band gaps for semiconductors are ~ 1 – 2 eV. Increasing temperature causes increased electrical conductivity by providing more carriers in the conduction band.

A variety of organic materials demonstrate reasonable conductivity, and a great deal of study has been devoted to conduction in organic materials. Initial study of conduction in organic materials was focused on molecular crystals (such as anthracene) that have weak intermolecular van der Waals interactions and low conductivities. However, the area of research has broadened and also includes charge-transfer compounds and polymeric materials. Research on new materials has led to synthesis of organic compounds with much larger conductivities than typical molecular crystals. Conducting and superconducting organic solids are now relatively commonplace. A listing of some compounds and associated conductivities are shown in Figure 2.

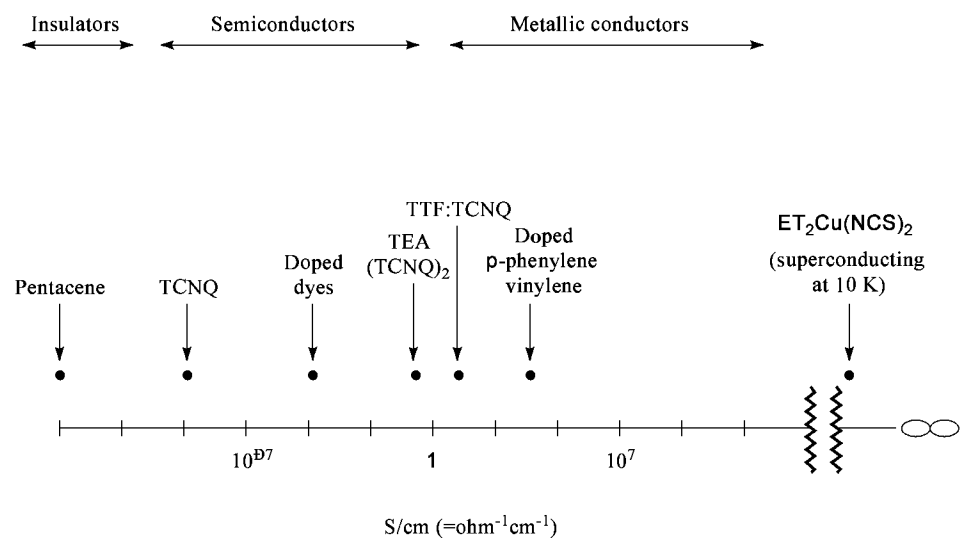


Fig. 2. Conductivity scale for different classes of organic materials.

Certain one-dimensional, chain-like organic semiconductors are found to display metallic properties (decreasing conductivity with increasing temperature) above a certain temperature (1). The phase transition that occurs in these stacked materials is termed a metal-to-insulator transition and is the result of the highly anisotropic nature of these materials. A stack of equally spaced atoms or molecules is subject to a reorganization to a lower symmetry configuration. This reorganization, termed a Peierls distortion (2), is the condensed matter counterpart of the molecular Jahn-Teller distortion. The Peierls deformation results in pairing of molecules along the chain with the formation of short and long intermolecular spacings. This nuclear deformation causes changes in the electronic distribution of the lattice. The result of this redistribution is the opening of an energy gap at the Fermi level, with resulting semiconductivity. Reviews of Peierls distortions can be found in the literature (3).

A common example of the Peierls distortion is the linear polyene, polyacetylene. A simple molecular orbital approach would predict *sp*² hybridization at each carbon and metallic behavior as a result of a half-filled delocalized π -orbital along the chain. Uniform bond lengths would be expected (as in benzene) as a result of the delocalization. However, a Peierls distortion leads to alternating single and double bonds (Fig. 3) and the opening up of a band gap. As a result, undoped polyacetylene is a semiconductor.

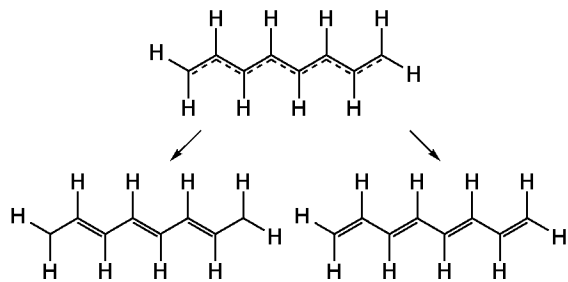


Fig. 3. Representation of the Peierls distortion in *trans*-polyethylene.

Interaction of molecules to form charge-transfer complexes occurs via transfer of electron density from one molecule (an electron donor, D) to another (an electron acceptor, A), resulting in a partially ionic ground state. The amount of charge transfer is large when the ionization potential of the donor is low and the electron affinity of the acceptor is high. Orbital descriptions and experimental aspects of charge-transfer complexes have been studied extensively, especially for weak complexes (4). In a charge-transfer salt, the intermolecular interactions give rise to periodic arrays of donors and acceptors.

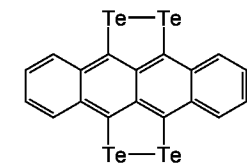
Typical charge-transfer salts form as stacks of planar D and A molecules, though the ratio of D:A need not be 1:1, as the interaction can be spread over more than two molecules. The amount of charge transfer (δ) per D ^{δ +} A ^{δ -} unit in the solid may be less than unity, with partial charges residing on the respective stacks. Interactions between stacks are usually weak, and the observed conductivity is, thus, anisotropic, with the preferred direction being along the stacking axis. Anion (A⁻) stacks result in electrical conductivity, while cation (D⁺) stacks conduct holes. Strongly bound charge-transfer systems are of primary interest for organic semiconducting applications. A listing of some common strong donors and acceptors is given in Table 1.

Table 1. Common Donor and Acceptor Molecules Used in Organic Semiconductor and Devices

Compound	CAS Registry Number	Structure
<i>Donors</i>		
2,2',5,5'-tetrathiafulvalene (TTF)	[31366-25-3]	
N,N,N',N'-tetramethyl- <i>p</i> -phenyl-enediamine (TMPD)	[100-22-1]	
5,6;11,12-tetratellurotetracene	[64479-92-1]	

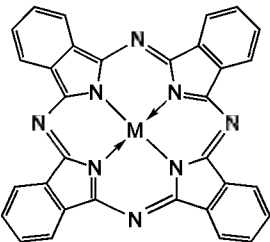
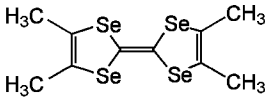
3,3',4,4'-tetramethyl-2,2',5,5'-tetraselenafulvalene (TMTSF)

[55259-49-9]



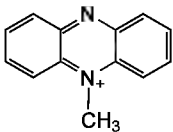
phthalocyanines (PC), M = Cu

[147-14-8]



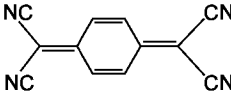
N-methylphenazinium (NMP)

[7432-06-6]



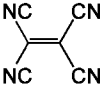
7,7,8,8-tetracyanoquinodimethane

Acceptors
[1518-16-7]



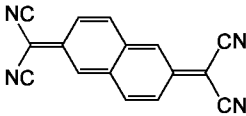
tetracyanoethylene (TCNE)

[670-54-2]



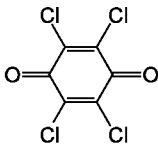
9,9,10,10-tetracyano-2,6-naph-thylenedimethane (TNAP)

[6251-01-4]



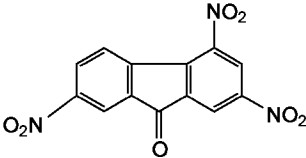
chloranil

[118-75-2]



trinitrofluorenone (TNF)

[129-79-3]



Some common single-carrier semiconducting salts are tetracyanoquinodimethane (TCNQ) radicals with alkali or other cations. The study of semiconductivity in these materials led to synthesis of more highly conductive charge-transfer salts based on the strong donor tetrathiafulvalene (TTF) with halogen or other acceptors. Organic materials have been developed with even larger conductivity and superconductivity above liquid helium temperatures.

Properties

Physical Properties: Electrical. Electrical properties have been the main focus of study of organic semiconductors, and conductivity studies on organic materials have led to the development of materials with extremely low resistivities and large anisotropies. A discussion of conductivity behaviors for various classes of compounds follows.

Charge-Transfer Salts. Most organic molecular crystals have only weak interactions between molecular units and have resulting low conductivities, with $S < 10^{-10} \Omega^{-1} \text{cm}^{-1}$. Charge-transfer solids tend to have higher conductivities than the molecular compounds from which they are composed, and semiconducting charge-transfer salts of TCNQ were first synthesized in the early 1960s. The simple metal-ion radical salts of the composition $M^+(TCNQ)^{\cdot -}$ have conductivities of $\sim 10^{-9} - 10^{-4} \Omega^{-1} \text{cm}^{-1}$ at room temperature, intermediate between conductivity values for insulators and typical metals. A variety of other systems show thermally activated conductivity in the range $10^{-10} - 1 \Omega^{-1} \text{cm}^{-1}$, including phthalocyanines and chloranil or tetracyanoethylene-based charge-transfer complexes. Conductivity in these semiconductive compounds obeys an Arrhenius expression:

$$\sigma = Ae^{-\Delta E/kT}$$

where $\mathcal{A}$ is a preexponential factor, k is Boltzmann's constant and ΔE is the activation energy for the thermally activated conductivity.

Proposed conductivity mechanisms and the importance of band motion vs hopping motion in semiconducting organic solids have generated a great deal of controversy. However, it is plausible that for many organic molecular semiconducting materials, the description of the conductivity lies somewhere between the extremes of localized hopping and delocalized band models (5).

The highly conductive class of solids based on TTF–TCNQ have less than complete charge transfer (~ 0.6 electrons unit for TTF–TCNQ) and display metallic behavior above a certain temperature. However, these solids undergo a metal-to-insulator transition and behave as organic semiconductors at lower temperatures. The change from a metallic to semiconducting state in these chain-like one-dimensional (1D) systems is a result of a Peierls instability. Although for true one-dimensional systems this transition should take place at 0 Kelvin, interchain interactions lead to effective non-1D behavior and inhibit the onset of the transition (6).

The temperature of the metal-to-insulator transition in TTF–TCNQ is 53 K. For systems with increased interchain coupling, the transition temperature for the onset of metallic conduction increases roughly as the square of the interaction between the chains. This behavior is true as long as the coupling between chains remains relatively weak. For compounds with strong interactions between stacks, the material loses its quasi-1D behavior. Thus, the Peierls distortion does not occur even at low temperatures, and the materials remain conductive.

The temperature range of metallic conductivity in organic materials can sometimes be increased by compaction of the compound under large pressures. For example, bis(tetramethyl tetraselenafulvalenium) hexafluorophosphate ((TMTSF)₂-PF₆) undergoes a metal-to-insulator transition at 12 K at atmospheric pressure. Under high pressures, however, the material stays conducting to low temperatures and goes superconducting at 0.9 K (7). A variety of organic-based materials have now been developed that have strong interchain interactions and possess superconductivity at low temperatures. Examples include bis(tetramethyl tetraselenafulvanenium) perchlorate (TMTSF)₂ClO₄ and compounds based on the donor molecule bis(ethylenedithio)tetrathiafulvalene (BEDT-TTF) with a variety of inorganic and organic anions (8).

Polymers. The use of polymers in semiconducting applications is growing rapidly. In the undoped form, most of these π -conjugated systems are semiconducting, with a small band gap separating filled valence and conduction bands. A listing of polymers commonly used in semiconducting applications is given in Table 2. The dark (nonphotoinduced) conductivity of these systems is typically low and increases with increasing temperature, as a result of thermalization of charge carriers. Doping the material leads to large increases in conductivity, with increases of several orders of magnitude possible. Doping removes electrons from the valence band or adds them to the conduction band. However, the charges are not delocalized along the polymer chain. An extra electron in the conduction band, for example, becomes localized as a result of a local geometry change caused by the excess charge. The result of the charge localization is termed a polaron, which as a result of the local energy minimum gives a state in the bandgap of the material. Conduction can then occur by motion of the polaron through the chain.

Table 2. Representative Polymers Used for Organic Semiconductors and Metals

Compound	CAS Registry Number	Structure
<i>Polymers</i>		
polyvinylcarbazole (PVK)	[25067-59-8]	
<i>trans</i> -polyacetylene	[25067-58-7]	
polythiophene	[25233-34-5]	
poly(<i>p</i> -phenylene)	[25190-62-9]	
polyaniline (PANI)	[25233-30-1]	
poly(<i>p</i> -phenylene vinylene) (PPV)	[26009-24-5]	

Conductivities of polymers of technological interest such as polypyrrole and polythiophene are typically $\sim 100\Omega^{-1}\text{cm}^{-1}$ in the doped state, and the conductivity can be tuned by reversibly doping and undoping the polymer. Derivatives of these and other polymers have achieved even higher conductivities.

One-dimensional polymers such as polyacetylene which have degenerate ground states are special cases. Conduction in polyacetylene has been extensively studied (9). The nature of conduction in pure, *trans*-polyacetylene is a result of the alternating single bond–double bond structure that results from the Peierls instability. The energy of the system is the same if the alternating pattern of the bonds is reversed. Conductivity has been described for this system in terms of solitons (mobile kinks in the polymer chain that link alternating bond patterns), excitations that are unique to systems with degenerate ground states (10,11). These neutral solitons are mobile, and extend over several bond lengths along the chain. As usually prepared, undoped polyacetylene has *p*-type conductivity as a result of acceptor impurities in the material. The conductivity increases with doping, reaching a plateau value at high doping levels (12). Highly doped *trans*-polyacetylene can display conductivities as high as $10^5\Omega^{-1}\text{cm}^{-1}$.

Optical Properties.

Charge-Transfer Salts. New strong absorbance bands (charge-transfer bands) are often observed in spectra of donor–acceptor solids. These bands are the result of excitation from a ground state that is predominantly donor in character to an excited state that resides mainly on the acceptor. For a given acceptor, the frequency of this band varies linearly (over a small range) with the ionization potential of the donor molecule. This absorption band can occur at lower energy than absorptions of the isolated D and A molecules, and is often in the infrared region. By a suitable choice of donor and acceptor, it is possible to tune the frequency of this absorption band (and the associated fluorescence) throughout the visible region. For compounds such as the simple charge-transfer salts of TCNQ (formula $M^+ TCNQ^-$) the charge transfer is complete, and these salts in solution exhibit spectra characteristic of the anion. In the solid state, an additional absorption band is observed as a result of charge transfer between adjacent TCNQ molecules: one acting as a π -donor and the other as a π -acceptor. Highly conducting salts such as those of the formula $M^+ TCNQ^-$ (TCNQ) have less than a complete negative charge per TCNQ molecule in the stack. These compounds exhibit an additional transition as a result of promotion of an electron from an anionic to neutral TCNQ site.

Polymers. Polymers used in semiconducting applications such as light-emitting diode displays are π -conjugated systems with small band gaps of 1–3.5 eV. In the undoped state they give strong π – π^* absorption bands in the visible or ultraviolet region, corresponding to transitions from the valence band to the conduction band levels. By tailoring the nature of the electron-donating and withdrawing substituents on the polymer backbone, absorbance and the resulting emission spectra can be shifted throughout the visible region. Conformational changes due to heating can drastically shift the absorption, suggesting possible uses as molecular switches.

Doping of the polymers brings about absorbances in addition to the main bandgap transition. These absorbances occur at lower energies, and grow in at the expense of the band-to-band transition. The new absorbances are due to transitions between states that lie in the band gap of the material (13). These midgap states are typically polaron states that arise due to the injected charge-carriers being localized by lattice distortions. As the doping levels are increased, absorption at lower energies continues to increase and the original band edge becomes obscured. At high doping levels the polymers show absorption throughout the ir and visible, consistent with metallic behavior.

Chemical Properties: Electrochemistry. Since charge-transfer and mobility of carriers are important features for organic semiconductors and devices, the energies of the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) are of interest. The energies of these levels are related to the efficiency of charge injection into a material as well as the degree of charge-transfer between donors and acceptors. Electrochemical measurements provide a convenient means for determining HOMO and LUMO levels by measuring the potentials required for oxidation and reduction of the material, respectively. Redox potentials also provide a quick method for checking energy gap values obtained from conductivity measurements.

Electron donor molecules are oxidized in solution easily. For example, $E_{1/2}^{ox}$ for TTF is 0.33V vs SCE in acetonitrile. Similarly, electron acceptors such as TCNQ are reduced easily. TCNQ exhibits a reduction wave at $E_{1/2}^{red} = 0.06V$ vs SCE in acetonitrile. The redox potentials can be adjusted by derivatizing the donor and acceptor molecules, and this tuning of HOMO and LUMO levels can be used to tailor charge-transfer and conductivity properties of the material. Knowledge of HOMO and LUMO levels can also be used to choose materials for efficient charge injection from metallic electrodes.

Reversible oxidation and reduction of polymers is commonly used to increase conductivity in these systems. Ions from the electrolyte are usually incorporated into the polymer as part of this process (see ELECTRICALLY CONDUCTING POLYMERS).

Stability.

Thermal Stability. The materials used in organic semiconducting applications are thermally labile upon exposure to high temperatures. For example, many of the compounds used in fabrication of organic light-emitting diodes (LEDs) are vapor deposited by resistive heating at relatively low temperatures in vacuum. Compounds such as the hole transporter *N,N'*-biphenyl-*N,N'*-bis(3-methylphenyl)-1,1' biphenyl-4,4' diamine (TPD) have glass-transition temperatures in the range of 150°C.

As a result of the organic nature of the materials, chemical oxidation is a problem when heating in atmosphere. For example, TTF melts at 119°C, with the formation of sulfoxides as a result of oxidation of the sulfur atoms in the molecule (14). Polymeric compounds can also oxidize at low temperatures. Polyacetylene is degraded at room temperature by atmospheric oxygen and water vapor. Although other conjugated polymers are typically more thermally stable, most exhibit degradation and decreased conductivity after heating in atmosphere.

Photostability and Atmospheric Stability. The presence of oxygen and water vapor has a profound effect upon the photostability of many organic semiconducting materials. For example, singlet oxygen has been found to be a highly reactive intermediate in the photooxidation of poly(phenylene vinylene) derivatives (15). Oxygen and water vapor have also been implicated in dark spot formation in the degradation of light-emitting diodes. As a result, performance lifetimes of these devices can be significantly prolonged by packaging under inert atmosphere. An exception to the general behavior of organic semiconducting systems is observed for *p-n* solar cells fabricated with the donor chloroaluminum phthalocyanine. These cells exhibit improved performance after exposure to water and oxygen, due to formation of a crystalline hydrate (16). This provides further evidence for the need for control over impurity and humidity levels during device production.

Structure

A discussion of structure in organic semiconductors includes chemical structure of individual molecules or unit cells, as well as three-dimensional structure of the bulk material. The structure of individual molecules can be tailored by synthetic methods, with molecules designed for specific applications. For example, rigid, planar structures are often used for dyes in photovoltaics, since this increases exciton mobility. Molecular packing and the structure of thin films can be dependent upon the method of fabrication and the structure of individual units. The nature of molecular and polymeric thin films for semiconducting applications ranges from highly crystalline to amorphous. Challenges ahead are to design better adhesion between organic films and substrates. Better interfacing and orientation are needed for many applications, since these can influence charge injection and transport. Though the structure of organic semiconductors is wide ranging, some common classes are discussed below.

Single-Stack Acceptor. Simple charge-transfer salts formed from the planar acceptor TCNQ have a stacked arrangement with the TCNQ units facing each other (intermolecular distances of ca 0.3 nm (~3). Complex salts of TCNQ such as $TEA(TCNQ)_2$ consist of stacks of parallel TCNQ molecules, with cation sites between the stacks (17). The interatomic distance between TCNQ units is not always uniform in these salts, and formation of TCNQ dimers (as in $TEA(TCNQ)_2$) and trimers (as in $Cs_3(TCNQ)_3$) can lead to complex crystal structures for the chainlike salts.

Single-Stack Donor. Ion-radical salts can also be formed from electron donors such as tetrathiafulvalene (TTF) or TMPD (*N,N,N',N'*-tetramethyl-*p*-phenylene diamine) with inorganic acceptors such as halogens. The resulting structure of compounds such as $TTF(A)_x$ (A = acceptor) is a linear chain of parallel stacked TTF molecules. The TTF molecules are registered directly above one another, with anions residing in sites between the TTF chains. Unlike the situation found for the TCNQ salts, the TTF molecules in these solids are equally spaced along the chain. Substantial π -overlap exists between the TTF molecules in the chain, providing for conduction along the stacking direction (Fig. 4).

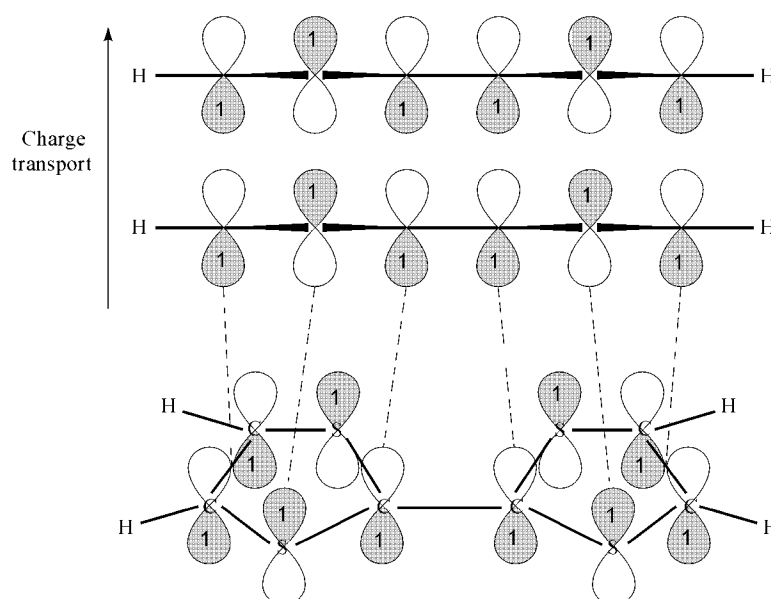


Fig. 4. Orbital overlap for conduction molecular orbitals and molecular stacking in tetrathiafulvalene (TTF).

A large family of compounds is based on combination of TTF-type donors and TCNQ-type acceptors. The resulting $D^{\delta+} A^{\delta-}$ solids form segregated stacks of D and A molecules with uniform spacings along the stacks. Molecules are tilted with respect to the stacking axis, and the π -overlap of the molecules on each chain results in the formation of bands and large conductivity along the stacking direction.

A number of other donor–acceptor molecular solids such as $\text{TMPD}^{\delta+} - \text{TCNQ}^{\delta-}$ or $\text{TMPD}^{\delta+} - \text{chloranil}^{\delta-}$ crystallize as mixed stacks of alternating D and A molecules. These compounds typically have much higher resistivities than the segregated salts because the alternating -DADA- sequence leaves no continuous channel for conduction.

Polymers. The individual units in a polymer are well-defined, but bulk polymer usually lacks the periodic structure characteristic of inorganic or small molecule crystalline solids. The structure of semiconducting and conducting polymers is determined by the units of the polymer backbone as well as pendant side-groups (18). Typical conducting polymers are conjugated systems that provide for motion of the electrons through the π -system, and both linear chain (poly-acetylenes) and aryl chain poly(*p*-phenylene) polymers are commonly used. As chain length increases during polymer growth, the long-chain polymers become insoluble and precipitate out during synthesis. The chain length in polymeric solids is thus, influenced by the nature of the solvent and the synthetic process. Although most polymers are fairly amorphous, chemical substitution has been used to design self-assembled oligomeric films with highly ordered structures (19). Film morphology for linear chain polymers can be affected by stretching of the film to orient the chains, resulting in higher crystallinity. The morphology of electrochemically grown conducting polymer thin films has also been shown to be dependent on the nature of the substrate (20).

The conducting polymer poly(sulfur nitride) is unusual in that it is crystalline, consisting of chains of sulfur and nitrogen packed in parallel. Individual chains are planar with alternating cis–trans conformation along the chain, and the interaction with neighboring chains leads to a highly ordered matrix.

Although its use as a transparent electrode in diodes has made it one of the most useful of the conductive polymers, the intractable nature of the material made the structure of polyaniline difficult to determine. Recent studies of polyphenyleneamineimines have conclusively shown that the structure of PANI is an exclusively para-linked system (21).

Synthesis and Manufacture

Donors. Common selenium and sulfur-containing donors such as tetrathiafulvalene [31366-25-3], tetramethyltetraselenafulvalene [55259-49-9], and bis-(ethylenedithio)tetrathiafulvalene [66946-47-3] are commercially available, as are other common donors such as tetramethyl-*p*-phenylenediamine [100-22-1]. Tellurium-containing donors can be prepared using a variety of synthetic routes. Tetratellurafulvalene, for example, is prepared in good yield using tin–lithium exchange (22) (Fig. 5).

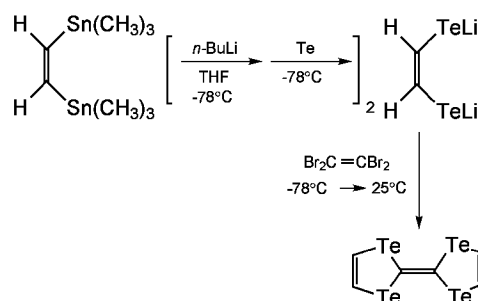


Fig. 5. Synthesis of the common donor tetratellurafulvalene.

Acceptors. Most common acceptor molecules such as tetracyanoethylene or tetracyanoquinodimethane are commercially available. However, TCNQ can be synthesized in high yield by a two-step synthesis involving a condensation of malonitrile with 1,4-cyclohexanedione followed by treatment with an oxidizing agent such as bromine or *N*-bromosuccinamide in pyridine solvent (23) (Fig. 6).

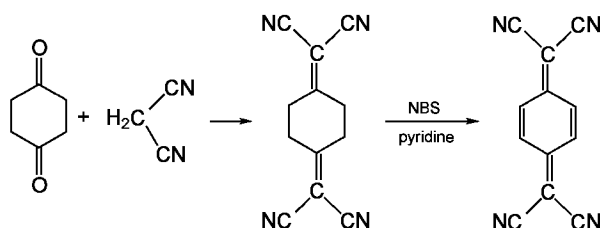


Fig. 6. Synthesis of the acceptor molecule tetracyanoquinodimethane (TCNQ).

Charge-Transfer Salts. Most charge-transfer salts can be prepared by direct mixing of donors and acceptors in solution. Semiconducting salts of TCNQ have been prepared with a variety of both organic and inorganic counterions. Simple salts of the type $M^+ TCNQ^-$ can be obtained by direct reaction of a metal such as copper or silver with TCNQ in solution. Solutions of metal iodides can be used in place of the metals, and precipitation of the TCNQ salt occur directly (24).

More highly conducting TCNQ salts of the form $M^+ (TCNQ)(TCNQ^-)$ are also easily synthesized, where M^+ can be any one of a variety of organic cations. These salts can be prepared by a number of different methods, though a general method involves mixing of a simple salt with TCNQ in solution. The precipitated salt can then be obtained by filtration.

Compounds containing the donor and acceptor moieties on the same molecule have also been prepared. A class of these compounds contains the D and A parts separated by sulfur atoms (a D- σ -A linkage), with the acceptor portion based on the TCNQ structure (25). Syntheses of these molecules can provide for the design of substituent groups to tailor packing and conducting properties in the solid state. Linkages of this D- σ -A type have garnered interest for possible use as unimolecular rectifiers (26).

Polymers. The π -conjugated polymers used in semiconducting applications are usually insulating, with semiconducting or metallic properties induced by doping (see ELECTRICALLY CONDUCTIVE POLYMERS). Most of the polymers of this type can be prepared by standard methods. The increasing use of polymers in devices in the last decade has led to a great deal of study to improve the processability of thin films of commonly used polymers.

Polyacetylene is the most studied of all of the polymers used in semiconducting applications. Early syntheses of polyacetylene were based on direct polymerization of acetylene using a Ziegler-Natta catalyst (usually titanium tetrabutoxide and triethylammonium) coated onto a glass substrate (27). The resulting polyacetylene films from these direct polymerizations were highly crystalline, insoluble, unstable with respect to atmosphere, and of low tensile strength. Other preparations from polymerization of acetylene have resulted in better film properties and higher conductivities; however, the films are not easily processible. To improve processibility and to generate new morphologies of polyacetylene that possess superior properties, precursor methods have been used (28). Polyacetylene has been prepared by ring-opening polymerization of polybenzvalene (29) as well as by polymerization involving addition of catalysts to 1,3,5,7 cyclooctatetraene (COT) (30). A variety of derivatives of polyacetylene can be generated with substituted precursor derivatives.

Polyaniline (PANI) can be formed by electrochemical oxidation of aniline in aqueous acid, or by polymerization of aniline using an aqueous solution of ammonium thiosulfate and hydrochloric acid. This polymer is finding increasing use as a "transparent electrode" in semiconducting devices. To improve processibility, a large number of substituted polyanilines have been prepared. The sulfonated form of PANI is water soluble, and can be prepared by treatment of PANI with fuming sulfuric acid (31). A variety of other soluble substituted *N*-alkylsulfonic acid self-doped derivatives have been synthesized that possess moderate conductivity and allow facile preparation of spincoated thin films (32).

Polythiophene can be synthesized by electrochemical polymerization or chemical oxidation of the monomer. A large number of substituted polythiophenes have been prepared, with the properties of the polymer depending on the nature of the substituent group. Oligomers of polythiophene such as (α -sexithienyl thiophene) can be prepared by oxidative linking of smaller thiophene units (33). These oligomers can be sublimed in vacuum to create polymer thin films for use in organic-based transistors.

Health and Safety Factors

Selenium is an essential element and is beneficial at low concentrations, serving as an antioxidant. Lack of selenium affects thyroid function, and selenium deficiencies have been linked to Keshan Disease (34). Selenium at high levels, however, is toxic. Hydrogen selenide (which is used in semiconductor manufacturing) is extremely toxic, affecting the mucous membranes and respiratory system. However, the toxicity of most organoselenium compounds used as donor compounds for organic semiconductors is not well studied.

Tellurium is not an essential element, and tellurium compounds are in general more toxic than their selenium counterparts. Metallic tellurium is known to have a teratogenic effect in rats, though no studies have been done on the toxicity of tellurium donor compounds (35).

The common acceptor molecule tetracyanoethylene is a poison, and sublimes at relatively low temperature (120°C). The toxicological effects of most cyano-type acceptors have not been fully investigated.

Uses

The performance of early organic semiconductor-based devices was far below that of their Si-based counterparts. In addition, the poor stability and reliability of many organic-based devices slowed their practical use. However, many problems have been overcome and these devices are becoming increasingly in a variety of applications. Organic semiconductors offer many advantages over inorganic-based materials for use in electronic and optoelectronic devices. Among the advantages are the possibility of tailoring properties through synthetic chemical modifications, the relative ease of generating large-area films, and low cost device manufacture. The following is a summary of some current uses and areas of research for potential applications.

Diodes. The first reports of electroluminescence (EL) from organic materials appeared in the 1960s (36–38), but it was not until 1987 that Tang and VanSlyke reported an Organic Light Emitting Diode (OLED) device with reasonable efficiency (39). The key to preparing an efficient device is to use a number of different metal–organic or organic thin-films materials, each serving a different function in the working device. Early devices consisted of two organic layers, one a tertiary amine and the other an aluminum coordination complex (aluminum-tris-8-hydroxyquinoline), the latter of which is responsible for bright green light emission. Over the last several years research on these novel devices has led to new organic and metal–organic materials with emission that covers the entire visible spectrum (40–46). Quantum efficiencies for these devices are comparable to LEDs based on inorganic semiconductors, typically ranging from 1–2.5% (photons/electrons) (41,47). OLEDs can achieve very high brightnesses, of greater than 15,000 cd/m² (43,48) and have operational lifetimes greater than 10,000 hours when driven at video brightness (100 cd/m²) (49).

OLEDs are based on amorphous (glassy) organic films, which can be deposited on virtually any substrate, ranging from rigid supports such as glass or silicon to highly flexible polymer supports (50). This is in contrast to inorganic LEDs, which require crystalline substrates for epitaxial growth of the active materials. An additional feature of OLEDs is that the organic layers are very thin (ca 50–100 nm, 500–1000 Å), and combined with large Franck-Condon shift between absorption and emission, makes them transparent to their own emission (51,52). This transparency may open the door to novel display architectures and applications. Taking these qualities together it is clear that OLEDs have unique characteristics as compared with other potential flat panel display (FPD) technologies. Although technical obstacles must be overcome, OLEDs have potential for certain display applications. This section is an overview of device structure and performance.

Device Structure and Operation. Efficient generation of electroluminescence in OLEDs depends on several factors. Injection of carriers (electrons and holes) from the respective electrodes, transport of carriers through the device, and radiative recombination of carriers must all be optimized. A simple OLED consisting of a thin organic or metal–organic film between an indium–tin–oxide (ITO) anode and a metal cathode is shown in Figure 7a. When a potential is applied to the device, the material is oxidized at the anode and reduced at the cathode, leading to the injection of holes and electrons, respectively, into the thin film. The holes and electrons injected into the thin film drift in the presence of the applied field via a hopping mechanism until they are removed at the opposite electrode or encounter an oppositely charged carrier within the film. In the latter case, electron–hole recombination results in the formation of an exciton and resultant emission of light.

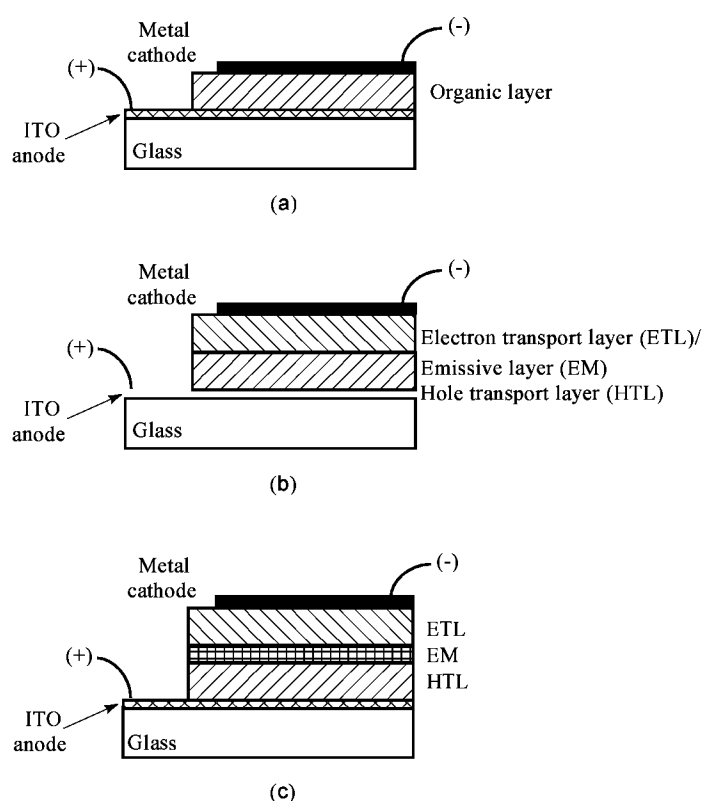


Fig. 7. Schematic of light emitting diodes: (a) single-layer device (b) single heterostructure (c) double heterostructure.

Single layer OLEDs have been fabricated with a variety of emitter molecules and conjugated polymers such as poly(phenylene vinylene) (PPV). These single-layer devices are typically not very efficient. To ensure high injection efficiency of both electrons and holes, multiple organic layers are used with each layer optimized for its particular role as a carrier injector or light emitter. The first efficient EL device was a single heterostructure (Fig. 7b), with a tertiary amine hole-transport layer (HTL) and an aluminum complex serving both as the electron-transport layer (ETL) and emissive material (EM) (39). In this device, the preferred conductivity of holes in the HTL and electrons in the ETL leads to a build up of carriers at the HTL–ETL interface. The excitons are formed at the interface, however, and hence are not spatially confined. As a result, it is important that the thickness of the ETL–EM layer emitter be chosen such that the majority of the excitons radiatively decay prior to reaching the adjacent electrode. The choice of ITO for the anode, and a suitable metal as the cathodes is based on the energies of the filled and vacant states of the organic materials relative to the work functions of these contacts. ITO is well matched to inject holes into the HTL HOMO, and low work function metals such as Mg, Al, and Ca can inject electrons into the ETL LUMO. Under forward bias (ITO positive) the injection of holes into the HTL and electrons into the ETL occurs at low (~ 5 V) potentials, while reversing the bias requires much higher potentials to inject the same current density. Hence, OLEDs show strongly rectifying current–voltage behavior.

An example "double heterostructure" OLED shown in Figure 7c uses an ITO coated glass substrate, upon which a hole transporting layer, typically composed of a tertiary amine (eg, *N,N'*-biphenyl-*N,N'*-bis(3-methylphenyl)1-1'-biphenyl-4,4'-diamine, abbreviated TPD), a thin film of an emissive material such as aluminum-8-hydroxyquinoline (Alq_3) and an electron-transporting layer (often an oxadiazole derivative) are sequentially deposited in vacuum (Fig. 8). This molecular multilayer is then capped with a cathode to complete the device. The HTL and ETL have high mobilities for holes and electrons, respectively, but very low mobilities for the oppositely charged carrier. When an electric field is applied to this device, holes will move through the HTL to the HTL–EM interface and electrons will move through the ETL to the ETL–EM interface. These carries then form a bound state exciton, and subsequently recombine in the EM, leading to efficient light emission. For efficient recombination at the emitter site, the emitter molecule should have a HOMO higher than that of the HTL and a LUMO lower than that of the ETL (53).

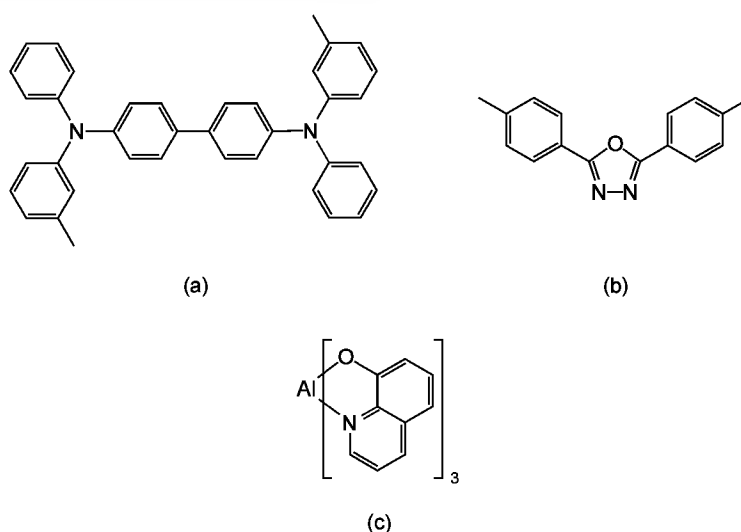


Fig. 8. Structures of (a) hole transporter molecule *N,N'*-biphenyl-*N,N'*-bis(3-methylphenyl)1-1'-biphenyl-4,4'-diamine (TPD); (b) an oxadiazole derivative; and (c) the common emitter, 8-hydroxyquinoline.

The picture presented above for confinement of the excitons within the device is for the EM layer sandwiched between the HTL and ETL. The EM need not be a discrete layer in the OLED, however, for exciton confinement to occur. Alternatively, the EM can consist of a luminescent molecule doped ($\sim 1\%$) into a polymeric or molecular host material (40,41,54,55). So long as the energy gap (or band gap) of the host is higher than that of the EM dopant, excitons will be effectively trapped or confined on the dopant molecules leading to improved EL efficiency. An example of such a dopant-based device

involves tetraphenylporphyrin doped into the Alq₃ layer of a TPD–Alq₃ OLED. Even at the dopant concentrations typically used (0.5–1%), the luminescence is entirely due to the tetraphenylporphyrin (54). A wide range of fluorescent and laser dyes have been developed with emission colors spanning the visible spectrum. These materials have very high luminescent quantum efficiencies in dilute solution, making them excellent candidates as emitter dopants in OLEDs, and indeed many have already been incorporated into efficient dopant-based OLEDs.

It has recently been shown that in some cases, the HTL and ETL materials do not need to be segregated into individual layers to achieve high device efficiency. For example, a homogeneous blend can be prepared with a polymeric hole transport (HT) material, and either a molecular or polymeric light emitter and electron transport (ET) material (53,56–61). An OLED prepared in this manner consists of a single organic film sandwiched between an ITO anode and a metal cathode. As in conventional devices, holes are injected from ITO and electrons from the metal cathode, with recombination occurring within the thin polymer film. Polymers and monomeric metal complexes have been used successfully in these blended devices to give colors that span the visible spectrum, with external quantum efficiencies >1% (56,57).

Emission. Generally, the shape of the electroluminescent spectrum is identical to that observed in photoluminescence. The reason for this can be seen by considering molecular LEDs. For closed shell molecules, the ground state S_0 has two electrons in the highest occupied molecular orbital (HOMO), and a completely vacant orbital for the lowest unoccupied molecular orbital (LUMO). In the S_1 excited state, both the HOMO and LUMO are singly occupied. Consider a molecular OLED with a thin organic film between an ITO anode and metal cathode. Since these films consist of weakly interacting molecules, the oxidized and reduced versions of the molecule correspond to the case where holes and electrons, respectively, are located at that molecular site. The holes and electrons move through the device, towards opposite electrodes. When the electron and hole combine, electron transfer from the reduced molecule (electron carrier) to the oxidized molecule (hole carrier) occurs, leading to a molecule to its ground electronic state, S_0 , and an adjacent one to the S_1 or Frenkel state (Fig. 9). The S_1 state then relaxes exactly as for photoluminescence. The close relationship between electroluminescence and photoluminescence is not surprising since the emission comes from the same excited state (or Frenkel exciton) in both processes. It should be noted that low levels of impurities can cause quenching of the emission or emission from the impurity due to exciton trapping. As a result, materials used in OLEDs are typically purified by repeated sublimation.

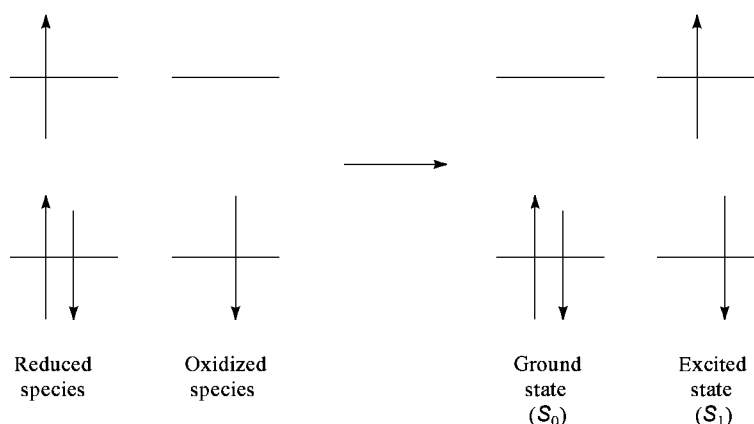


Fig. 9. Representation of exciton formation in molecular LEDs.

Single emitter devices can be made in a variety of colors, with good efficiency. The color of the device emission can be altered in several ways. One of the most common ways is the use of dopants, as mentioned earlier. First, energy transfer from the host to the dopant is maximized when the overlap of the host emission spectrum and the dopant absorption spectrum is favorable. By the use of suitable dopants, emission can be tuned and narrowed. In both molecular and polymeric OLEDs, synthetic methods can be used to tailor the emission. For example, replacing the CH in the 4 position of the quinolate ligand of Alq with a heterocyclic N leads to a 60 nm red shift in luminescence relative to Alq (62). Narrowing and shifting of the emission can also be effected through resonant cavity design. By tailoring the thickness and dielectric constants of the layers, the profile of broad emission from a device can be improved (63).

Device Stability. Stability of OLEDs has been a significant problem, since many organic materials are susceptible to oxidation. The mechanism of degradation in PPV-based devices has been shown to be related to the formation of singlet oxygen (64). Similarly, metal quinolate devices degrade quickly upon exposure to atmosphere water vapor and oxygen. Packaging devices under anaerobic conditions has been shown to extend device lifetime considerably, and devices with operating lifetimes of over 10,000 hours have been reported (49). The role of impurities in device degradation has not been fully established, and the behavior of the organic–contact interface remains an area of intense study. The thermal properties of the organic materials also limit their ultimate lifetime. Increasing the glass-transition temperatures without lowering the carrier mobilities of organic materials is a very active area of research, which may lead to more stable devices or devices that can function well at elevated temperatures.

Efficiency. Efficiency of a device can be reported in terms of an internal quantum efficiency (photons generated / electrons injected). The external quantum efficiency often reported is lower, since this counts only those photons that escape the device. Typically only a fraction of photons escape, due to refraction and waveguiding of light at the glass interface (65). The external efficiency can be increased through the use of shaped substrates (60).

Regardless of device design, for maximum quantum efficiency it is important to have a highly fluorescent emissive center. Many emitters are therefore highly absorbing and rigid, qualities that lead to large fluorescence. Because of spin selection rules governing recombination of the electron and hole, only one quarter of recombinations will lead to singlet excited states. The rest will result in triplet states, which are usually not efficient. This limits the internal quantum efficiency to a nominal maximum of 25%. However, other processes such as triplet–triplet annihilation may increase this value, and the question of the maximum possible efficiency is still not resolved. Though efficiencies for devices have climbed significantly in recent years, further improvements in efficiency are possible.

Outlook. OLEDs in their current form meet the requirements for several low resolution applications, and with improvements in reliability, efficiency and color tuning they should soon find use in a number of display applications. One area of application will be flexible displays. Organic OLED devices (both molecular and polymer based) on polymer supports have remarkable stability even after repeated bending at sharp angles (67,68). This high degree of substrate flexibility and device stability make these polymer-supported devices attractive for a range of applications, including portable roll-up displays and conformable displays for attachment to windows, windshields, or instrument panels. Aside from bendable devices, transparent OLEDs (TOLEDs) can be created by using a semitransparent metal cathode (69). These devices are >71% transparent when turned off, with light emitted from both top and bottom device surfaces when voltage is applied.

There are technical problems to be overcome before OLEDs can become competitive with liquid crystal technology for use in full color, high definition displays such as computer screens. Improvements in both color saturation and reliability are needed. A review of the prospects of OLEDs for a variety of uses has been published (70). Improvements in color saturation may be realized through the use of microcavities, with layer design used to select red, green, or blue emission from a single broad-emitter material (71,72). Recent development of stacked, independently addressable OLEDs may also lead to improvements in resolution, with three devices (R,G,B) in a vertical configuration providing minimal pixel size (73). Further improvements in the technology should lead to increasing use of OLEDs in a variety of applications.

Photovoltaic Devices. For many inorganic semiconductors, absorption of light can be used to create free electrons and holes. In an organic semiconducting solid, however, absorption of a photon leads to the formation of a bound electron–hole pair. Separation of this pair in an electric field can

occur due to differing mobilities of the electron and hole, and this separation of the charge particles gives rise to a photovoltage (see PHOTOVOLTAIC CELLS). This photovoltaic effect is of great interest for use in solar cells, since the conversion of light energy from the sun to electrical energy provides a nonpolluting renewable energy source. The photovoltaic effect is also of interest for photodetectors and for memory storage devices. Photocurrents are of great importance in biological systems as well, and there is interest in light-harvesting devices based on protein pigments (74).

Solar cells and other photovoltaic devices have typically been based on silicon or other inorganic materials, which can provide high power conversion efficiencies (>25%). However, organic materials can also be used to generate a photovoltage. Photovoltaic effects have been studied in a wide variety of materials including solvent-cast films of Buckminsterfullerene (C_{60}) (75) and self-assembled films of alkyl-derivatized metaloporphyrins (76). Many of these systems produce a photovoltage; however, the efficiency is lower than that found in inorganic-based devices.

Device Structure. In both inorganic and organic materials, absorption of light generates electrons and holes, but these will recombine and lead to a loss in device efficiency unless they can be separated and made to flow through an external circuit. There are many types of junctions that are used to separate the electron and hole. In single-crystal inorganic devices, a $p-n$ junction is created by adjoining precisely doped p and n -type regions. For organic devices, a $p-n$ junction can be fabricated by sequential deposition of thin films of n -type (electron accepting) and p -type (electron donating) organic materials. Schottky junctions, formed at semiconductor-metal interfaces, are often used in simple organic photovoltaic cells. In all cases, the potential across the junction is used to sweep the photogenerated carriers toward the opposite electrodes.

A typical organic-based photovoltaic cell (Fig. 10) consists of thin films of an organic semiconducting material sandwiched between conductive surfaces, at least one of which must be transparent (such as glass-indium tin oxide). The organic layer is thin, typically ca 100 nm thick, and can be deposited by solvent evaporation or by sublimation under vacuum (77). The organic layer must be relatively thin to minimize resistance in the device. It must also be uniform and defect free (since defects can serve as traps for the charge carriers) and must be also free of pinholes which would lead to short circuiting of the device. In order to maximize efficiency, highly absorbing species are often incorporated into the device. Unlike inorganic devices, organic photovoltaic cells offer the possibility of modifying device properties such as spectral sensitivity through changing of chemical substituents of molecules in the device.

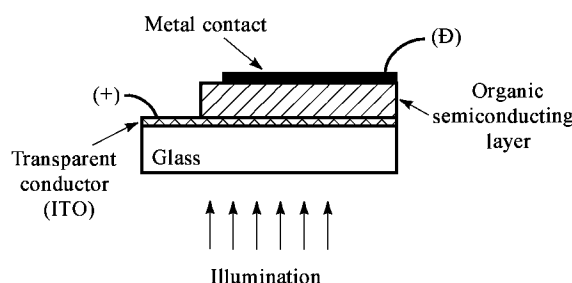


Fig. 10. Typical design for single-layer photovoltaic device.

Device Performance. Early sandwich type devices using metal-free phthalocyanines (molar absorptivities of $\sim 10^5$) between gold electrodes were reported to have quantum efficiencies of 0.16% (power conversion efficiencies of 0.009%) (78). Single-dye, Schottky-type sandwich cells have been fabricated from both n -type dyes such as rhodamine B and p -type dyes such as merocyanine (79) with appropriate metal contacts. Although single-layer cells have been developed with improved efficiency compared to early devices, cells based on $p-n$ heterojunctions promise superior performance. These cells possess advantages over single-dye devices, including a broader spectral range. Heterojunction solar cells typically consist of two films of organic dyes between indium tin oxide (ITO) and metallic electrodes. To facilitate charge separation, the dyes have large differences in Fermi levels. Highly efficient organic based cells have been fabricated utilizing thermally robust phthalocyanines and perylene derivatives (16). These $p-n$ type devices provide absorption throughout the visible spectrum and monochromatic internal quantum efficiencies of up to 20%.

Aside from dye-based cells, a good deal of study has been devoted to photovoltaic effects in polymer-based devices. Single-layer sandwich-type devices composed of aluminum-polyphenylenevinylene-ITO have been fabricated with solution-cast polymer layers of $\sim 0.5\mu$. These devices display quantum efficiencies of up to 5% (power efficiencies of 0.07%) (80), though the efficiency decreases at high illumination levels (80). Other single-layer polymer devices fabricated with polyacetylene show similar efficiencies (81). To obtain optimum efficiency in a photovoltaic device, it is necessary to dissociate the exciton before the charges recombine to generate a photon, and a resulting loss in device efficiency. Multilayer devices consisting of separate donor and acceptor layers generate substantial improvement in device efficiency by facilitating dissociation of the electron-hole pair. In these two-layer devices, excitons that diffuse to the D-A interface are separated (the electron going to the acceptor), and the charges can then be swept on to the electrodes through their respective layers.

However, not all excitons have sufficiently long lifetimes to reach the interface before recombining. To circumvent this problem and increase device efficiency, heterostructure devices have been fabricated. In these devices, donors and acceptors are mixed together to create a network that provides many internal interfaces where charge separation can occur. Heterostructure devices made from the donor polymer poly(2-methoxy-5-(2'-ethyl)-hexyloxy- p -phenylenevinylene) (MEH-PPV) mixed with the acceptor polymer CN-PPV (a cyano-substituted derivative of PPV) showed improved performance (82).

Outlook. Organic materials could provide much cheaper solar cells, since crystalline or epitaxially grown inorganic materials are costly. Also, deposition of organic compounds in vacuum could provide for large-area or custom shaped cells. Recent advances have greatly improved the efficiency of polymer-based solar cells, though they still suffer from a lower efficiency than inorganic-based devices. Also, the polymers suffer from photostability problems that must be overcome. However, they still could see near-term use in applications which do not require exposure to powerful light sources.

Transistors. The use of molecular materials for transistors is attractive because of the facile processability and low cost of organic thin films. Early organic thin-film transistors (TFTs) based on lutetium bis-phthalocyanine (83), polyacetylene (84), or polythiophene derivatives (85) as the active component were shown to have source-drain currents that could be modulated with applied gate voltage. However, performance of early devices was poor compared to inorganic-based devices, and turn-on times were slow. Mobilities for phthalocyanine devices of $\mu = 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and $\mu = 10^{-5} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ for polythiophene devices are some four to five orders of magnitude below the mobilities for Si-based devices. Though stable field-effect transistors based on metal phthalocyanines have been developed with mobilities of $10^{-3} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, much of the research on organic field-effect transistors has focused on semiconducting polymeric materials. Early devices typically generated the polymer electrochemically under inert atmosphere (86) or by spincoating; devices with enhanced performance were fabricated using vacuum deposition of oligomers. Devices from α -conjugated oligomers of polythiophene (α -sexithienyl, α -6T) showed mobility enhancement after heat treatment, with maximum $\mu = 10^{-3} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ reported (87). These p -type polycrystalline thin films (thicknesses of 100–200 nm) give on/off current ratios of 10^2 – 10^3 .

"All-organic" transistors based on α -6T were designed with organic materials serving as the semiconductor, insulator, and support (the source and drain contacts were metallic) (88). Aside from being flexible as a result of the organic support, these devices showed very high mobilities. The reported field-effect mobility (4.6×10^{-1}) and transit times are similar to those of amorphous-silicon based transistors. The performance of these devices was noted to be highly dependent on the nature of the semiconductor-insulator interface. Later refinement of α -6T based devices has improved the ratio of source-drain currents in the on and off states, with ratios of $>10^6$ obtained (89). Aside from the p -type polymer, hole transporting devices, n -type devices have been fabricated. In addition, devices have been developed that can function in either n or p type modes. Heterostructure devices consisting of C_{60} and α -6T active layers provide hole or electron drain currents depending upon the gate bias (90). The improvements in device performance are continuing, and organic transistors are likely to become important in a number of applications ranging from displays to integrated circuits.

Reproduction and Resist Materials. Organic photoreceptors have advantages over inorganic-based systems in cost and flexibility. Organic charge-transfer complexes, polymers, and dyes have found wide use in xerography (see ELECTROPHOTOGRAPHY). The role of the photoreceptor is to convert an optical input signal to an electrostatic map on a photoconductive surface. The photoreceptor must accept and hold a surface charge, discharging only in those areas that are exposed to illumination. Oppositely charged toner particles are then attracted to the undischarged areas of the photoconductor to create an image representing the dark regions of the original image.

The photoreceptor must be highly optically absorbing throughout the visible region, must be stable with respect to light and chemicals such as ozone, and must be relatively insulating in the absence of illumination. The field of organic photoconducting materials has been reviewed (91,92).

The first use of organic charge transfer complexes was of poly(*N*-vinylcarbazole) [25067-59-8] (PVK) mixed with the acceptor 2,4,7-trinitro-9-fluorenone [129-79-3] (TNF) in a 1:1 ratio (93). In this device, charge generation through absorption and charge transport both occur in a single layer. This material had good sensitivity throughout the visible region, but the carcinogenic properties of the fluorenone led to the discontinuation of its commercial use. A number of other acceptors have been used as dopants in PVK, including tetracyanoethylene, tetracyanoquinodimethane, and chloranil (94). Charge-transfer complexes from triarylamine donors and boron diketone acceptors have also been investigated for use as photoreceptors when doped into a Lexan host (95). The use of single-layer devices has been supplanted by two-layer systems, with one layer used for charge generation and the other for charge transport. Highly absorbing organic dye molecules are typically usually used for charge-generation and doped polymers serve as transport layers.

Batteries. Many π -conjugated polymers can be reversibly oxidized or reduced. This has led to interest in these materials for charge-storage batteries, since polymers are lightweight compared to metallic electrodes and liquid electrolytes. Research on polymer batteries has focused on the use of polymers as both the electrode and electrolyte. Typical polymer electrolytes are formed from complexes between metal-ion salts and polar polymers such as poly(ethyleneoxide). The conductivity is low at room temperature due to the low mobility of cations through the polymer-matrix, and the batteries work more efficiently when heated above the glass-transition temperature of the polymer. Advances in the development of polymer electrolytes have included polymers poly(ethylene oxide) intercalated into layered silicates (96). These solid-phase electrolytes exhibit significantly improved conductance at room temperature.

A variety of polymers have been investigated for use as electrodes, including polyacetylene and polyaniline. The stability of polymer electrodes with respect to repeated cycling has been a question. Although polyacetylene degrades after a limited number of cycles under oxidation, its performance for reduction is much better (97). Polymers for electrode applications must not only be robust with respect to air and repeated cycling, but must be easily processable, and cheap, and must provide high power-weight ratios. One efficient rechargeable system uses a thidiazole-based polymer which undergoes depolymerization as the battery is discharged, with the electrode being regenerated upon recharging (98). Improvements in polymer battery design should make polymer-based cells attractive, especially for low power applications.

Electrochromics. Oxidation or reduction of organic compounds often results in a color change (see ELECTROCHROMICS). Materials exhibiting such changes are termed electrochromics. Typically, the organic is pale in color prior to oxidation or reduction, with intense color change upon application of a voltage as a result of new optical absorption bands. Electrochromic materials are of interest for use in displays as well as for optical shutters. An important parameter in performance of these devices is the read-write efficiency, which is the percentage of electrically generated color that can be removed by reversing the current through the device.

The divalent cation methyl viologen is highly colored in reduced form and is used as an electrochromic in solution cells (Fig. 11).

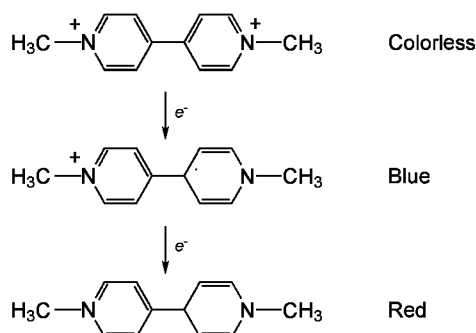


Fig. 11. Electrochromic behavior of methyl viologen.

Other solution-based electrochromic systems have been based on TCNQ, which is soluble in organic solvents such as acetonitrile. One-electron reduction forms the highly colored TCNQ⁻ anion, which absorbs throughout the visible region. Polymeric derivatives of TCNQ have been chemically attached to electrode surfaces to generate reversible electrochromic systems (99). Tethering the polymer to the electrode surface prevents diffusion of species away from the electrode surface and results in improved read-write efficiencies. Other organic acceptors such as *o*-chloranil (3,4,5,6-tetrachlorobenzoquinone) have also been used in electrochromic cells, demonstrating excellent stability and highly colored radical forms.

Tetrathiafulvalene (TTF) has also been used in electrochromic devices. TTF-based polymers spin-coated onto transparent electrode surfaces form stable thin films with reproducible electrochromic properties (100). The slow response of these devices has been attributed to the rate of ion movement through the polymer matrix.

The use of polymeric materials for electrochromic displays is especially promising. A number of conjugated polymer systems can be used as electrochromics, with insulating and doped forms of the polymer exhibiting different absorption spectra (101,102). Electrochemical oxidation of aniline results in thin films of polyaniline, which can exist in various redox states. Successive oxidations cause reversible color changes from transparent to yellow-green to deep blue-black (103). Polythiophene can also be electrochemically generated from the monomer. Though polythiophene electrochromic films suffer a loss in read-write efficiency after a large number of read-write cycles, derivatives of polythiophene have found use as electrochromes in a number of devices including display (104) and memory (105) devices. The colors in the oxidized and reduced forms of these systems depend upon the nature of the substituents on the monomer unit.

Switches and Sensors. The use of conducting polymers has been shown to be promising for chemical sensor technology, with molecular recognition leading to changes in conductivity of the polymer. One approach utilizes polythiophene derivatives which undergo ion-specific conformational changes, resulting in conductivity reductions upon sensing (106). Switching phenomena in charge-transfer complexes has also been observed. Thin films of metal-organic charge-transfer complexes such as Cu-TCNQ can be switched between low and high conductivity states with application of a suitable field. Conduction in the high conductivity state has been associated with channel formation in the material (107). One design for optoelectronic gates incorporates a dye molecule as the absorbing chromophore with a series of linked metalloporphyrins serving as a molecular wire. Switching is observed as a result of the change in the porphyrin energy levels (and thus the conduction in the wire) upon reversible oxidation at a designated site (108). Polymeric systems also show promise as temperature sensors, since thin films of doped polymers such as polyacetylene exhibit temperature dependent color changes.

BIBLIOGRAPHY

"Semiconductors, Organic," in *ECT* 3rd ed., Vol. 20, pp. 674-698, by Robert C. Haddon, Martin L. Kaplan, Fred Wudl, Bell Laboratories.

1. F. Gutman, H. Keyzer, L. Lyons, and R. B. Samoano, *Organic Semiconductors*, part B, Robert E. Kreiger Publishing Co., Malabar, Fla., 1983, pp.

- 399–412.
2. Peierls, *Quantum Theory of Solids*, Oxford University Press, 1972.
3. M. H. Whangbo, in J. S. Miller, ed., *Extended Linear Chain Compounds*, Plenum Press, New York, 1982, p. 127.
4. R. S. Mulliken and W. B. Person, *Molecular Complexes*, John Wiley & Sons, Inc., New York, 1969.
5. Z. Soos, *J. Chem. Ed.* **55**(9), 549 (1978).
6. Ref. 1, p. 400.
7. D. J  rome, A. Mazaud, M. Ribault, and K. Bechgaard, *J. Phys. Lett (Orsay)* **41**, L95 (1980).
8. U. Geiser, J. A. Schluter, H. H. Wang, A. M. Kini, J. M. Williams, P. P. She, H. I. Zakowicz, M. L. VanZile, and J. D. Dudek, *J. Am. Chem. Soc.* **118**, 9996 (1996).
9. J. Kanicki, in T. A. Skotheim, ed., *Handbook of Conducting Polymers*, Vol. 1, Marcel Dekker, New York, 1986, pp. 544–660.
10. W. P. Su and co-workers, *Phys. Rev. Lett.* **42**, 1698 (1979).
11. A. O. Patil, A. J. Heeger, and F. Wudl, *Chem. Rev.* **88**, 183 (1988).
12. J. Kanicki, S. Bou  , and E. Vander Donckt, *Thin Solid Films* **92**, 243 (1982).
13. H. Eckhardt, L. W. Shacklette, K. Y. Jen, and R. L. Elsenbaumer, *J. Chem. Phys.* **91**(2), 1303 (1989).
14. K. Bechgaard, in W. E. Hatfield, ed., *Molecular Metals*, Plenum Press, New York, 1979, p. 1.
15. B. H. Cumpston, K. F. Jensen, F. Klavetter, E. G. J. Staring, and R. C. J. E. Demandt, *Mat. Res. Soc. Proc.* **413**, 35 (1996).
16. J. B. Whitlock, P. Panayotatos, G. D. Sharma, M. D. Cox, R. R. Sauers, and G. B. Bird, *Optical Eng.* **32**(8), 1921 (1993).
17. Ref. 1, p. 242.
18. H. R. Allcock and F. W. Lampe, *Contemporary Polymer Chemistry*, 2nd ed., Prentice-Hall, Englewood Cliffs, N.J., 1990.
19. F. Garnier, A. Yassar, R. Hajlaoui, G. Horowitz, F. Deloffre, B. Servet, S. Ries, and P. Alnot, *J. Am. Chem. Soc.* **115**, 8716 (1993).
20. I. Rubinstein, J. Ripshon, E. Sabatani, A. Redondo, and S. Gottesfeld, *J. Am. Chem. Soc.* **112**, 6315 (1990).
21. F. Wudl, R. O. Angus, Jr., F. L. Lu, P. M. Allemand, D. J. Vachon, M. Nowak, Z. X. Liu, and A. J. Heeger, *J. Am. Chem. Soc.* **109**, 3677 (1987).
22. R. D. McCullough, M. D. Mays, A. B. Bailey and D. O. Cowan, *Synth. Met.* **27**, B487–B491 (1988).
23. D. S. Acker and W. R. Hertler, *J. Am. Chem. Soc.* **84**, 3370 (1962).
24. L. R. Melby, R. J. Harder, W. R. Hertler, W. Mahler, P. E. Benson, and W. E. Mochel, *J. Am. Chem. Soc.* **84**, 3374 (1962).
25. P. Bando, N. Martin, J. Segura, C. Seoane, E. Orti, P. Viruela, and R. Viruela, *J. Org. Chem.* **59**, 4618–4629 (1994).
26. R. M. Metzger, in R. R. Birge, ed., *Molecular and Biomolecular Electronics*, American Chemical Society, Washington, D.C., 1994, p. 91.
27. T. Ito, H. Shirakawa, and S. Ikeda, *J. Polym. Sci., Polym. Chem. Ed.* **12**, 11 (1974).
28. J. H. Edwards and W. J. Feast, *Polym. J.* **21**, 595 (1980).
29. T. M. Swager and R. H. Grubbs, *J. Am. Chem. Soc.* **111**, 4413–4422 (1989).
30. F. L. Klavetter and R. H. Grubbs, *J. Am. Chem. Soc.* **110**, 7807 (1988).
31. J. Yue and A. J. Epstein, *J. Am. Chem. Soc.* **112**, 2800–2801 (1990).
32. S. Chen and G. Hwand, *J. Am. Chem. Soc.* **117**, 10055–10062 (1995).
33. D. Fichou, G. Horowitz, B. Xu, and F. Garnier, *Synth. Met.* **89**, 243 (1990).
34. M. Corn, ed., *Handbook of Hazardous Materials*, Academic Press, 1993 p. 650.
35. M. Sittig, ed., *Handbook of Toxic and Hazardous Chemicals and Carcinogens*, 2nd ed., Noyes Publications, Park Ridge, N.J., 1985.
36. J. Dresner, *RCA Rev.* **30**, 322 (1969).
37. W. Helfrich and W. G. Schneidere, *Phys. Rev. Lett.* **14**, 229 (1965).
38. D. F. Williams and M. Schadt, *Proc. IEEE* **58**, 476 (1970).
39. C. W. Tang and S. A. VanSlyke, *Appl. Phys. Lett.* **51**, 913 (1987).
40. C. W. Tang, S. A. VanSlyke, and C. A. Chen, *J. Appl. Phys.* **65**, 3610 (1989).
41. C. Hosokawa, H. Higashi, H. Nakamura, and T. Kusumoto, *Appl. Phys. Lett.* **67**, 3853–3855 (1995).
42. J. Kido and K. Nagai, *JALCOM* **192**, 30–33 (1993).
43. Y. Hamada, T. Sano, M. Fujita, T. Fujii, Y. Nishio, and K. Shibata, *Jpn. J. Appl. Phys.* **32**, 514–515 (1993).
44. M. Takeuchi, H. Masui, I. Kikuma, M. Masui, T. Muranoi, and T. Wada, *Jpn. J. Appl. Phys.* **31**, 498–500 (1992).
45. C. Adachi, T. Tsutsui, S. Saito, *Appl. Phys. Lett.* **56**, 799–801 (1990).
46. C. Hosokawa, H. Higashi, H. Nakamura, and T. Kusumoto, *Appl. Phys. Lett.* **67**, 3853–3855 (1995).
47. P. E. Burrows, Z. Shen, V. Bulovic, D. M. McCarty, S. R. Forrest, J. A. Cronin, and M. E. Thompson, *J. Appl. Phys.* **79**, 91–8005 (1996).
48. Y. Hamada, T. Sano, M. Fujita, T. Fujii, Y. Nishio, and K. Shibata, *Chem. Lett.*, 905–906 (1993).
49. S. A. VanSlyke, C. H. Chen, and C. W. Tang, *Appl. Phys. Lett.* **69**, 2160 (1996).
50. G. Gustafsson, G. M. Treacy, Y. Cao, F. Klavetter, N. Colaneri, and A. J. Heeger, *Synth. Met.* **57**, 4123 (1993).
51. V. Bulovic, G. Gu, P. E. Burrows, S. R. Forrest, and M. E. Thompson, *Nature* **380**, 29 (1996).
52. V. Bulovic, G. Gu, P. E. Burrows, S. R. Forrest, and M. E. Thompson, *Appl. Phys. Lett.* **68**, 2606–2608 (1996).
53. G. E. Johnson, K. M. McGrane, and M. Stolka, *Pure Appl. Chem.* **67**, 175–182 (1995).
54. P. E. Burrows, S. R. Forrest, S. P. Sibley, and M. E. Thompson, *Appl. Phys. Lett.* **69**, 2959 (1996).
55. J. Kido, M. Kimura, and K. Kagai, *Science* **267**, 1332 (1995).
56. J. Tian, C. C. Wu, M. E. Thompson, J. C. Sturm, R. A. Register, M. J. Marsella, and T. M. Swager, *Adv. Mater.* **7**, 395–398 (1995).
57. J. Pommerehne, H. Vestweber, W. Guss, R. F. Mahrt, H. Bassler, M. Porsch, and J. Daub, *Adv. Mater.* **7**, 551–554 (1995).
58. C. Zhang, H. von Seggern, B. Kraabel, H.-W. Schmidt, and A. J. Heeger, *Synth. Met.* **72**, 185–188 (1995).
59. J. Kido, M. Kohda, K. Okuyama, and K. Nagai, *Appl. Phys. Lett.* **61**, 761–763 (1992).
60. C. Zhang, H. von Seggern, K. Pakbaz, B. Kraabel, H.-W. Schmidt, and A. J. Heeger, *Synth. Met.* **62**, 35–40 (1994).
61. J. Kido, H. Shionoya, and K. Magai, *Appl. Phys. Lett.* **67**, 2281–2283 (1995).
62. Z. Shen, P. Burrows, D. Z. Garguzov, M. McCarty, M. E. Thompson, and S. R. Forrest, *J. Appl. Phys.* **35**, L401 (1996).
63. N. Takada, T. Tsutsui, and S. Saito, *Appl. Phys. Lett.* **63**, 2022 (1993).
64. R. D. Scudlock, B. Wang, P. R. Ogilby, J. R. Sheats, and R. L. Clough, *J. Am. Chem. Soc.* **117**, 10194–10202 (1995).
65. N. C. Greenham, R. H. Friend, and D. C. Bradley, *Adv. Mater.* **6**, 491 (1994).
66. G. Gu, D. Z. Garbazov, P. E. Burrows, S. Venkatesh, S. R. Forrest, and M. E. Thompson, *Optics Letters* **22**, 397 (1997).
67. G. Gustafsson, G. M. Treacy, Y. Cao, F. Klavetter, N. Colaneri, and A. J. Heeger, *Synth. Met.* **55–57**, 4123 (1993).
68. G. Gu, P. E. Burrows, S. Venkatesh, S. R. Forrest, and M. E. Thompson, *Optics Letters* **22**, 172 (1997).
69. V. Bulovic, G. Gu, P. E. Burrows, S. R. Forrest, and M. E. Thompson, *Nature* **380**, 29 (1996).
70. L. J. Rothberg and A. J. Lovinger, *J. Mater. Res.* **11**(12), 3174 (1996).
71. A. Dodabalapur, L. J. Rothberg, T. M. Miller, and E. W. Kwock, *Appl. Phys. Lett.* **64**, 19 (1994).
72. A. Dodabalapur, L. J. Rothberg, and T. M. Miller, *Electronics Lett.* **30**, 1000 (1994).
73. Z. Shen, P. E. Burrows, V. Bulovic, S. R. Forrest, and M. E. Thompson, *Science* **276**, 2009 (1997).
74. F. T. Hong, in Ref. 26, p. 554.
75. C. H. Lee, G. Yu, D. Moses, A. J. Heeger, and V. I. Srdanov, *Appl. Phys. Lett.* **65**, 6, 664 (1994).
76. C.-Y. Liu, H.-L. Pan, H. Tang, M. A. Fox, and A. J. Bard, *J. Phys. Chem.* **99**, 7632 (1995).

77. R. N. Marks, R. Zamboni, and C. Taliani, *Mat. Res. Soc. Symp. Proc.* **247**, 814 (1996).
78. K. J. Hall, J. S. Bonham, and L. E. Lyons, *Aust. J. Chem.* **31**, 1661 (1978).
79. D. Wühde and D. Meissner, *Adv. Mat.* **3**, 129 (1991).
80. H. Antoniadis, B. R. Hsieh, M. A. Abkowitz, S. A. Jenekhe, and M. Stolka, *Synth. Met.* **62**, 265–271 (1994).
81. J. Kanicki, in T. A. Skotheim, ed., *Handbook of Conducting Polymers*, Vol. 1, Marcel Dekker, New York, 1986, p. 543.
82. J. J. M. Halls, C. A. Walsh, N. C. Greenham, E. A. Marseglia, R. H. Friend, S. C. Moratti, and A. B. Holmes, *Nature* **376**, 498 (1995).
83. R. Madru, G. Guillaud, M. Al Sadoun, M. Maitrot, J.-J. Andriü, J. Simon, and R. Even, *Chem. Phys. Lett.* **145**(4), 343 (1988).
84. F. Ebisawa, T. Kurosawa, and S. Nara, *J. Appl. Phys.* **54**, 3255 (1983).
85. A. Assadi, C. Svensson, M. Willander, and O. Inganäs, *Appl. Phys. Lett.* **53**, 195 (1988).
86. A. Tsumura, H. Koezuka, and T. Ando, *Synth. Met.* **25**, 11 (1988).
87. G. Horowitz, D. Fichou, X. Peng, Z. Xu, and F. Garnier, *Solid State Comm.* **72**, 381 (1989).
88. F. Garnier, G. Horowitz, X. Peng, and D. Fichou, *Adv. Mat.* **2**, 592 (1990).
89. A. Dadabalapur, L. Torsi, and H. E. Katz, *Science* **268**, 270 (1995).
90. A. Dodabalapur, H. E. Katz, L. Torsi, and R. C. Haddon, *Science* **269**, 1560 (1995).
91. P. M. Borsenberger and D. S. Weiss, in A. S. Diamond, ed., *Handbook of Imaging Materials*, Marcel-Dekker, New York, 1991, p. 379.
92. P. M. Borsenberger and D. S. Weiss, *Organic Photoreceptors for Imaging Systems*, Marcel Dekker, New York, 1993.
93. W. D. Gill, *J. Appl. Phys.* **43**, 5033 (1972).
94. H. Meier, W. Albrecht, E. Zimmerhackl, N. Geheeb, and U. Tschirwitz, *Polym. Bull.* **7**, 505 (1982).
95. T. E. Goliber and J. H. Perlstein, *J. Chem. Phys.* **80**, 4162 (1984).
96. S. Wong, S. Vasudevan, R. A. Vaia, E. P. Gianelis, and D. B. Zax, *J. Am. Chem. Soc.* **117**, 7568 (1995).
97. R. B. Kaner and A. G. McDiarmid, *Synth. Met.* **14**, 3 (1986).
98. C. Vaughan, *New Scientist*, **31**, 36 (1990).
99. R. W. Day, G. Inzelt, J. F. Kinstle, and J. Q. Chambers, *J. Am. Chem. Soc.* **104**, 6804 (1982).
100. F. B. Kaufman, A. H. Schroeder, E. M. Engler, S. R. Kramer, and J. Q. Chambers, *J. Am. Chem. Soc.* **102**, 483 (1990).
101. M. Gazard, in T. A. Skotheim, ed., *Handbook of Conducting Polymers*, Marcel Dekker, New York, 1986, Chapt. 19.
102. P. M. S. Monk, R. J. Mortimer, and D. R. Rosseinski, *Electrochromism: Fundamentals and Applications*, VCH Publishing Co., Weinheim, 1984.
103. T. Kobayashi, H. Yonehama, and H. Tamura, *J. Electroanal. Chem.* **161**, 419 (1984).
104. E. M. Genies, M. Lapowski, C. Santier, and E. Viel, *Synth. Met.* **18**, 631 (1987).
105. K. Yoshino, R. Sugimoto, J. G. Rose, and W. F. Schmidt, *Jpn. J. App. Phys.* **24**, L33 (1985).
106. T. M. Swager, M. J. Marsella, Q. Zhou, and R. J. Newland, *Mat. Res. Soc. Symp. Proc.* **413**, 407 (1996).
107. H. Duan, M. D. Mays, D. O. Cowan, and J. Kruger, *Synth. Met.* **28**, C675–680 (1989).
108. R. W. Wagner, J. S. Lindsey, J. Seth, V. Palaniappan, and D. F. Bocian, *J. Am. Chem. Soc.* **118**, 3996 (1996).

Scott P. Sibley
Goucher College

SMART MATERIALS

On the front page of Section A of the January 16, 1997 issue of *USA Today* was a cover story titles "SmartStory: How "smart" became marketing's hot new buzzword" (1). The article listed numerous examples of smart materials and structures. Several examples of these included Simply Smart Home, Smart Saw, Smart Chopper, Smart Clapper, Smartwidgets, and Smart Dog (all trademarks). During the first nine months of 1996, 925 trademark applications were filed containing the word smart in its name. Only 35 smart-containing trademark applications were filed in 1985. There are over 3415 active trademark files containing smart in the name. For comparative purposes, trademarks that use "dumb", "dummy", "dummies", or "dum" total 133. Things trademarked smart outnumber things trademarked pretty 12 to 1, and big almost 1.25 to 1.

What are smart materials? If one uses the *USA Today* as a guide, one could simply say that a smart material is one that attracts the high technological oriented consumer. However, from a technical and simple point of view, a smart material is a material that responses to its environments in a timely manner (2–5). To expand on this definition, a smart material is one that receives, transmits, or processes a stimulus and responds by producing a useful effect, which may include a signal that the material is acting upon it. Stimuli may include strain, stress, temperature, chemicals, an electric field, a magnetic field, hydrostatic pressures, different types of radiation, and other forms of stimuli. Transmission or processing of the stimulus may be in the form of an absorption of a photon, of a chemical reaction, of an integration of a series of events, of a translation or rotation of segments within the molecular structure, of a creation and motion of crystallographic defects or other localized conformations, of an alteration of localized stress and strain fields, and of others. The useful effects produced could be a color change, an index of refraction change, a stress or strain distribution change, or a volume change. Also, incorporated within the definition of smart materials is the ability to be reversible.

Under the proper set of environments and circumstances all materials are smart and depict smart behavior at some point during their life cycle. Some examples of technically smart behaving materials are piezoelectric materials, electrostrictive materials, magnetostrictive materials, electrotheological materials, magnetorheological materials, thermoresponsive materials, pH-sensitive materials, uv-sensitive materials, smart polymers, smart gels (hydrogels), smart catalysts, and shape-memory alloys. Technical examples of smart (intelligent) structures and applications include structures embedded with fiber optics such as composite-materials, actuators, sensors, microelectromechanical systems (MEMS), vibration control, sound control, shape control, product health or lifetime monitoring, cure monitoring, intelligent processing, active and passive controls, self-repair (healing), artificial organs, novel indicating devices, designed magnets, damping, aeroelastic stability, and stress distributions (6). These applications perform in systems ranging from space systems, fixed-wing and rotary-wing aircraft, automotive, civil structures, machine tools, recreation, and medical devices.

Smart structures are structures that incorporate at least one smart material within itself and from the effort produced by the smart material causes an action. A term that is implied in the description of a smart structure is *biomimetic*. A smart structure has sensors (nerves), actuators (muscles), and a control system (brain). This article deals with the materials themselves. The structures will be covered in more detail in the upcoming *Encyclopedia of Smart Materials* to be published by John Wiley & Sons. Searching various databases yields hundreds of citations per year since 1992 and if one search is combined with another, it is surprising how many citations are not common to each other. The number of societies (Society of Photo-Optical Instrumentation Engineers, The Materials Research Society, American Chemical Society, etc) giving symposiums on smart structures has increased both in number and frequency. If the Web sites on the Internet are included, the amount of information is mind boggling. The Web sites discuss the embedding of sensors in various type of structures, the different types of sensors such as fiber optics, gas, chemical, biological, and others that are available, the types stimuli these systems detect, the types of responses given, the types of control features that they offer, and the design of the smart structures. There has been a vast growth of the implementation of fiber optics (7,8) (qv), conductive polymers (9–11), and nonlinear optical materials (12,13) (qv).

With the inception of smart materials and structures, the world is on the brink of a new material age. Within this century the polymer age and the composite age have been experienced. These have now been followed by the smart material age. The implementation of smart materials into smart structures describes the smart material age more accurately.

Piezoelectric Materials

Piezoelectric materials are materials that exhibit a linear relationship between electric and mechanical variables. Electric polarization is proportional to mechanical stress. The direct piezoelectric effect can be described as the ability of materials to convert mechanical stress into an electric field, and the reverse, to convert an electric field into a mechanical stress. The use of the piezoelectric effect in sensors is based on the latter property. For materials to exhibit the piezoelectric effect, the materials must be anisotropic and electrically poled; ie, there must be a spontaneous electric field maintained in a particular direction throughout the material. A key feature of a piezoelectric material involves this spontaneous electric field and its disappearance above the Curie point. Only solids without a center of symmetry show this piezoelectric effect, a third-rank tensor property (14,15).

There are two principal types of materials that can function as piezoelectrics: the ceramics and polymers. The piezoelectric materials most widely used are the piezoceramics based on the lead zirconate titanate, PZT, formations, mixed sodium and potassium niobates, lithium niobate, and quartz. The advantages of these piezoceramics are that they have a high piezoelectric activity and they can be fabricated in many different shapes.

A new class of materials called smart tagged composites has been developed for structural health monitoring applications. These composites consist of PZT-5A particles embedded into the matrix resin (unsaturated polyester) of the composite (16).

Polyimide films with metal-filled conductive particles have been suggested for numerous types of microelectronics-related applications and in hybrid microelectronic technologies. Some of these materials have not reached their full potential owing to processing difficulties. An interesting study involved a conductive graphite-filled polyimide that was compatible with conventional microelectronics processing techniques (17). The characterization of this graphite-filled polyimide showed a quadratic dependence of the resistance on the deflection. The piezoeresistive coefficient was at a maximum at 18° graphite loading, the residual stress was unaffected by variations in graphite loading, and Young's modulus increased with high graphite loading. Traits such as those, made this material amenable for microsensor or actuator uses.

Piezoelectric materials are used in many different types of sensing–actuating devices. A few applications include printing, monitoring of performance behavior of adhesive joints, and intelligent processing.

Electrostrictive Materials

Electrostrictive materials are materials that exhibit a quadratic relationship between mechanical stress and the square of the electric polarization (14,15). Electrostriction can occur in any material. Whenever an electric field is applied, the induced charges attract each other, thus, causing a compressive force. This attraction is independent of the sign of the electric field and can be approximated by

s = M Z^2 + Q P^2

The strain *s* lies along the axis of the electric field, *E*, or most often along the axis of the induced polarization, *P*. The electrostrictive coefficients for the electric field and polarization are *M* and *Q*, respectively. Electrostriction is a small effect. In contrast to piezoelectric materials, electrostrictive materials

show a large effect near the Curie temperature, especially for ferroelectric substances such as members of the perovskite family. Ferroelectrics are ferroic solids that have domain walls that can be moved by external forces or fields. There are three principal types of ferrois solids (ferro-electrics, -magnetics, and -elastics).

Typical electrostrictive materials include such compounds as lead manganese niobate:lead titanate (PMN:PT) and lead lanthanum zirconate titanate (PLZT). Electrostriction is a fourth-rank tensor property observed in both centric and acentric insulators (14,15).

Magnetostrictive Materials

As materials show mechanical deformation induced by electric fields, the same type of material response can be observed when the stimuli is a magnetic field. Shape changes are the largest in ferromagnetic and ferrimagnetic solids. The repositioning of domain walls that occur when these solids are placed in magnetic field leads to hysteresis between magnetization and an applied magnetic field. All of these effects disappear when the ferromagnetic material is heated above its Curie temperature. Ferrimagnetic materials have macroscopic properties similar to ferromagnetics; however, their microscopic properties are different. The magnetic dipoles of a ferromagnetic solid are aligned parallel to each other, whereas in a ferrimagnetic the alignment can be either parallel or in other directions (14,15).

Materials that have shown a response to a magnetic stimuli have primarily been inorganic in chemical composition, alloys of iron, nickel, and cobalt doped with rare earths. TERFENOL-D, an alloy of terbium, dysprosium, and iron, Tb_xDy_{1-x}Fe_y, with *x* between 0.27 and 0.30 and *y* between 1.90 and 1.95, is probably the most effective magnetostrictive material. The name TERFENOL is an acronym for two of the elements present in the alloy and NOL refers to the Naval Ordinance Laboratory where this type of material behavior was developed. Magnetostriction occurs at its fullest potential in crystalline materials. Cost still appears to be one of the hindrances of magnetostrictives in reaching commercial importance. However, over the past three decades there has been a great interest in the development of organic and organometallic magnets. Several key features of these new magnets due to their organic nature include that these types of magnets do not have extended three-dimensional structural networks and they can be made using common solvents at low temperatures with a multitude of different synthetic techniques.

Magnets that exhibit commercial potential must have properties that are based on their critical temperature, limiting (saturation) magnetization, and coercive field. Magnets are useful below their critical temperature. For most inorganic type magnets, their critical temperatures are substantially above room temperature. The first organometallic-based magnet, the ionic salt complex of ferric bis(pentamethylcyclopentadienide) and tetracyanoethylene, [Fe(C₅(CH₃)₅)]⁺ [(NC)₂C≡C(CN)₂]⁻, is a ferromagnet below its critical temperature of 4.8 K. The highest effective critical temperature of an organometallic magnet is approximately 400 K (127°C) (18).

In comparing organic magnets with organometallic magnets there are several key differences between the two types. The first is obvious; the organic-based magnets do not contain metal atoms. The discovery of the first organic-based magnet (or magnets without metals or metal salts) caused a great deal of rethinking the principles of magnetism. The second difference involves the fact that in organic-based magnets, the coupled spins reside entirely in the *p* orbitals; whereas in the organometallic-based magnets, they are either in the *p* or *d* orbitals, or a combination of the two. A series of over a thirteen organic-based magnets based upon nitroxide chemistry was synthesized that showed magnetic behavior between 0.6 K and 1.48 K. The most studied compound of this series was 4-nitrophenylnitronyl nitroxide. It showed a saturation magnetization corresponding to one spin per molecular, thus, indicating ferromagnetic behavior. The nitroxide, 1,3,5,7-tetramethyl-2,6-diaza-adamantane-*N,N'*-dioxyl, showed the highest critical temperature of that group studied (18).

To illustrate the significance of the organic portion of these magnets whenever the counteranions in [Fe(C₅(CH₃)₅)]⁺ [(NC)₂C≡C(CN)₂]⁻ was changed to 7,7,8,8-tetracyano-*p*-quinodimethane, TCNQ, the resulting material became metamagnetic (18). The alignment of electron spins in this class of materials changes when they are exposed to a strong external magnetic field (over 1600 oersted below a critical temperature of 2.55 K). Ferrous chloride belongs to the metamagnet family of materials. The [Fe(C₅(CH₃)₅)]⁺ [TCNQ]⁻ complex was unique because it was the first member of the metamagnets to exist without extending covalent bonding. Also, the changing of TCNE acceptor to TCNQ increased the spin density in the acceptor portion of the salt, thus, enhancing spin coupling and stabilizing the ferromagnetic behavior of the material. In other studies, bulkier substituent groups were added to the cyclopentadiene ring and other metal ions were substituted in place of iron to enhance the magnetic properties of these organometallic-based magnets.

Electrorheological Materials

Electrorheological materials (fluids) have been know for over a hundred years. An electrorheological (ER) fluid, is one whose viscous properties are modified by applying an electric field (14,15). There are a wide assortment of electrorheological fluids, which are usually a uniform dispersion or suspension of particles within a fluid. A typical example of an electrotheological fluid is a mixture of corn starch in a silicone oil. The mechanism of how electrorheological fluids work is simple. In an applied electric field the particles orient themselves in fiber-like structures (fibrils). When the electric field is off, the fibrils disorient themselves. The damping characteristics of the system can be changed (flexible to rigid). A means of viewing how ER fluids work can be found in a study of the use of electrorheological fluids as a smart space material. Electrorheological fluids were investigated using a single-link flexible-beam test bed. The beam was a sandwich configuration with ER fluids distributed along its length. When the beam is rapidly moved back and forth, the ER fluid is used to provide flexibility during the transient response period of the maneuver (for speed) and made rigid at the end point of the maneuver (for stability). A more practical means of viewing ER fluids is to compared their actions to fly fishing.

Electrorheological fluids are non-Newtonian fluids, that is, the relationship between shear stress and shear strain rate is nonlinear. The changes in viscous properties of electrorheological fluids are only obtained at relatively high electric fields in the order of 1 kV/mm. The practical applications of electrorheological fluids center around their abilities to transferring shear stresses and of acting as a variable damping material in an electric field.

Magnetorheological Materials

Magnetorheological materials (fluids) are the magnetic equivalent of electrorheological fluids. In this case, the particles are either ferromagnetic or ferrimagnetic solids that are either dispersed or suspended within a liquid and the applied field is magnetic (14).

An interesting adaptation of magnetorheological fluids is a series of elastomeric matrix composites embedded with magnetic particles such as iron. During the thermal cure of the elastomer, a strong magnetic field was applied to align the iron particles into chains. These chains of iron particles were locked into place within the composite through the cross-linked structure of the cured elastomer. The resistance of the composite to changes in modulus or deformation was controlled by an external magnetic field. When stimulated by a compressive force, the composite was 60% more resistant to deformation in a magnetic field. When subjected to a shear force, the magnetic-field induced modulus of the composite was an order of magnitude higher in comparison to its modulus measured in a zero magnetic field (19).

Thermoresponsive Materials

Polymeric materials are unique owing to the presence of a glass-transition temperature. At the glass-transition temperatures, the specific volume of the material and its rate of change changes, thus, affecting a multitude of physical properties. Numerous types of devices could be developed based on this type of stimuli-response behavior; however, this technology is beyond the scope of this article.

Materials that typify thermoresponsive behavior are polyethylene-poly(ethylene glycol) copolymers that are used to functionalize the surfaces of polyethylene films (smart surfaces) (20). When the copolymer is immersed in water, the poly(ethylene glycol) functionalities at the surfaces have solvation behavior similar to poly(ethylene glycol) itself. The ability to design a smart surface in these cases is based on the observed behavior of inverse temperature-dependent solubility of poly(alkene oxide)s in water. The behavior is used to produce surface-modified polymers that reversibly change their hydrophilicity and solvation with changes in temperatures. Similar behaviors have been observed as a function of changes in pH (21-24).

Other examples of materials that respond smartly to changes in temperature are the poly(ethylene glycol)s-modified cottons, polyesters, and

polyamide–polyurethanes fabrics. Coupling the thermoresponsiveness of these fabrics with a sensitivity to moisture resulted in a family of fabrics that have properties of thermal adaptability and reversible shrinkage. An interesting application of these materials is as smart pressure bandages. These fabrics, when exposed to an aqueous medium such as blood, contract, thus applying pressure to the wound to stop the bleeding. As long as the fabric is wet, it applies pressure; once dry, the fabric releases the pressure (19).

pH-Sensitive Materials

By far the most widely known classes of pH-sensitive materials are the classes of chemical compounds that include the acids, bases, and indicators. The most interesting of these are the indicators. These materials change colors as a function of pH and usually are totally reversible (see HYDROGEN ION CONCENTRATION).

In addition to acting as a means of observing changes in pH in titrations and in chemical reactions, indicators have been used in the development of novel chemical indicating devices. During the 1970s, a series of these devices were developed based on the permeability of organic vapors through polymeric films or through a porous polymeric plug. The devices contained a double-polymer-film were used. One film was to control the diffusion of the organic vapor and the second was a microporous film to protect the system when it was activated. The rest of the device consisted of a paper strip or a porous plug impregnated with a base or an acid (the opposite chemistry as what was in the vial) and an indicator. The system is activated by crushing the vial. The volatile component permeates through the film, it wets and reacts with the chemistry on the paper or in the plug, and in the presence of an indicator, a color change is observed. At higher temperatures, the color change proceeds faster, and at lower temperatures, slower. A family of devices was developed with different time and temperature profiles. These devices were developed to monitor the shipments of time and or temperature-sensitive items such as pharmaceuticals, foods, other perishables, and tasks, such as the changing of intravenous fluids in a hospital.

Other examples of pH-sensitive materials such as the smart hydrogels and smart polymers have been included in other sections of this article.

Light-Sensitive Materials

There are several different types of material families that exhibit different types of response to a light stimuli (14,15). One type is materials that exhibit electrochromism. This behavior is typically a change in color as a function of an electrical field. Other types of behaviors include thermochromism (color change with heat), photochromic materials (reversible light-sensitive materials), photographic materials (irreversible light-sensitive materials), and photostritive materials (shape changes due to light usually caused by changes in electronic structure).

An interesting material with both electro- and thermochromism behavior, Li_xVO_2 , was evaluated for a "smart window" application (25). Films of Li_xVO_2 were prepared by reactive sputtering and annealing an electrolyte of LiClO_4 and propylene carbonate.

Materials are being developed to exhibit both photochromic and photographic behaviors; one such system is based on a substituted indolinospiro-benzopyrene embedded in a polystyrene matrix (26). This system acts as a photochromic system at low exposure in the uv range and at high exposure it functions as a photographic system. The image can be devisualized by heat and can be restored many times with uv irradiation.

Smart Polymers

Even though smart polymers have been used in all types of applications and can exhibit all types of stimuli–response behaviors, the term *smart polymers* has been used as a separate category describing smart materials. The distinction in smartness between materials and polymers can be confusing at times. In medicine and biotechnology, smart polymer systems usually pertain to aqueous polymer solutions, interfaces, and hydrogels. They refer to polymeric systems that are capable of responding strongly to slight changes in the external medium; a first-order transition, accompanied by a sharp decrease in the specific volume of the system. The presence of a poor solvent is one of the main conditions for this phenomenon in swollen polymer networks or linear polymers to occur. A poor solvent causes the forces of attraction between the polymer chain segments to overcome the repulsion forces associated with the extended volume, thus leading to the collapse of the polymer chain (30). Owing to the recent flurry of activities with smart gels (hydrogels), they are treated separately in this article.

Smart polymers can response to environmental stimuli such as temperature, pH, ions, solvents, reactants, light or uv radiation, stress, recognition, electric fields, and magnetic fields. These stimuli once acted upon, result in changes in phases, shape, optics, mechanics, electric fields, surface energies, recognition, reaction rates, and permeation rates. The polymers that fit into this category include the natural occurring polymers, acrylic polymers and copolymers, and polymers based on combining acid monomers with basic monomers (27,28).

In addition to the polymer systems mentioned elsewhere, one polymer that has been used for a great number of smart applications is poly(vinylidene fluoride) (PVDF). The interesting features of this polymer include its structure, morphology, extraordinarily high piezoelectric and pyroelectric activities, and optical properties (29,30). A smart application of PVDF is in the area of smart aperture antennas. A recent study developed a series of antennas capable of variable directivity and power density. The concept of actuating these antennas was based on a voltage drop across a film structure of PVDF bonded to metallized Mylar. The change in voltage across the PVDF–Mylar caused the material to expand or contract. This movement caused a moment to develop within the structure, thus, in turn, causing a change in shape (31).

Smart Gels (Hydrogels)

Smart (intelligent) gels (or hydrogels) do not illustrate new technology. They are finally reaching commercialization after thirty years of research and development. The concept of smart gels is also more complex than the simple concept of solvent-swollen polymer networks. It is the behavior of the solvent-swollen polymer networks in conjunction with the material being able to response to other types of stimuli; such as temperature, pH, and concentrations of solvents (32–34). The phenomenon was first observed in swollen clear polyacrylamide gels. Upon cooling, the gels would cloud up and became opaque. Warming the gels would regain its clarity. Further investigations into this type of behavior showed that some systems could expand to hundreds of times its volume or could collapse to expel up to 90% of its fluid content with a stimulus of a 1°C change in temperature. These types of behaviors have lead to the recent development of gel-based actuators, valves, sensors, controlled-released systems for drugs and other substances, artificial muscles for robotic devices, chemical memories, optical shutters, molecular separation systems, and of course, toys. Other potential systems for development with smart gels or smart hydrogels include paints, protective coatings, adhesives, recyclable absorbents, bioreactors, bioassay systems, and display devices.

An example of a smart gel chemical composition consists of an entangled network of two polymers; one is a poly(acrylic acid) (PAA), and the second, a triblock copolymer containing poly(propylene oxide) (PPO) and poly(ethylene oxide) (PEO) in a PEO–PPO–PEO sequence. The PPA portion of the smart gel system is a bioadhesive and is pH responsive. The PPO segments are hydrophobic that help solubilize lipophilic substances in medical applications and the PPO segments tend to aggregate, thus resulting in gelation at body temperatures. Other types of polymers involved in smart gels include citosan, a hydrolyzed derivative of chitin (a polymer of *N*-acetylglucosamine which is found in shrimp and crab shells), a copolymer of poly(*N*-isopropylacrylamide) and poly(acrylic acid), and a graft copolymer of poly(methacrylic acid) and poly(ethylene glycol). These smart gels are being developed for controlled delivery of insulin in diabetic patients and they can be custom-engineered to be temperature, pH, and glucose responsive. Other smart gels are being tailored to be responsive to changes in ionic strength, pressure, stress, light intensity, electric fields, and magnetic fields. Smart hydrogels containing acid groups swell in a weakly alkaline medium, but collapse in an acid medium. On the other hand, smart hydrogels with basic functional groups behave in an opposite manner; ie, swell in an acidic medium and collapse in a basic medium. Polyampholytic smart hydrogels swell to the maximum extent at neutral pH values. When such gels are subjected to either acidic or basic media they undergo rapid dehydration (35). These smart hydrogens are copolymers of methacrylic acid and 2-(*N,N*-dimethylamino)ethyl methacrylate.

Smart Catalysts

The development of smart catalysts is a relatively new field of investigation. One class of smart catalysts is based on homogeneous rhodium-based poly(alkene oxide)s, in particular those with a poly(ethylene oxide) backbone. Traditionally chemical-catalyzed reactions proceed in a manner in which the catalysts becomes more soluble and active as the temperature is raised. This can lead to exothermal runaways, thus, posing both safety and yield problems. These smart catalysts behave differently. As the temperature increases, they become less soluble, thus, precipitating out of solution and inactive. As the reaction mixture cools down, a smart catalysts redissolves and becomes active again (19). Other smart catalysts are being developed that dissociates at high temperatures (less active) and recombines at low temperatures (more active) (36).

Smart Memory Alloys

Some materials undergo a thermomechanical changes as it passes from one phase to another. The crystalline structure of materials such as alloys based on nickel and titanium enters the martensitic phase as the alloy is cooled below a critical temperature. In this stage, the alloy is easily manipulated through large strains with a little change in stress. As the temperature of the alloy is increased above the critical (transformation) temperature it changes into the austentic phase. In the austentic phase, the alloy regains its high strength and high modulus. It behaves like a "normal" metal. The alloy shrinks during the transformation from the martensitic to austentic phase (14,15).

The use of shape-memory alloys as actuators depends on their use in the plastic martensitic phase that has been constrained within the structural device. As stress is generated within the shape memory alloy, it is transmitted to the device, thus, triggering an action. The behavior of shape memory alloys is very complex. The material changes its modulus coupled with a change in mechanical losses, acts as a variable damping material in its martensite phase, and has a large energy storing capacity. The properties of shape memory alloys are further complicated by issues of the chemical composition of the alloys and past histories.

Shape memory alloys (SMAs) can be divided into three functional groups: one-way SMAs, two-way SMAs, and magnetically controlled SMAs (37). The magnetically controlled SMAs show great potential as actuator materials for smart structures because they could provide rapid strokes with large amplitudes under precise control. The most extensively used conventional shape memory alloys are the nickel–titanium- and copper-based alloys. Due to their low cost, iron-based shape memory alloys are becoming more popular in smart structure applications. Iron–manganese–silicon steels, iron–manganese–silicon steels alloyed with chromium, nickel, and cobalt, and iron–manganese–silicon steels alloyed with nitrogen all fit into this category.

As previously mentioned, the nickel–titanium alloys have been the most widely used shape memory alloys. This family of nickel–titanium alloys is known as Nitinol (Nickel Titanium Naval Ordnance Laboratory in honor of the place where this material behavior was first observed). Nitinol have been used for military, medical, safety, and robotics applications. Specific usages include hydraulic lines capable of F-14 fighter planes, medical tweezers, anchors for attaching tendons to bones, eyeglass frames, underwire brassieres, and antiscalding valves used in water faucets and shower heads (38,39). Nitinol can be used in robotics actuators and micromanipulators that simulate human muscle motion. The ability of Nitinol to exert a smooth, controlled force when activated is a mass advantage of this material family (5).

Elastorestrictive Materials

This class of smart materials is the mechanical equivalent of electrostrictive and magnetostriictive materials. Elastorestrictive materials exhibit high hysteresis between strain and stress (14,15). This hysteresis can be caused by motion of ferroelastic domain walls. This behavior is more complicated and complex near a martensitic phase transformation. At this transformation, both crystal structural changes induced by mechanical stress and by domain wall motion occur. Martensitic shape memory alloys have broad, diffuse phase transformations and coexisting high and low temperature phases. The domain wall movements disappear with fully transformation to the high temperature austentic (paraelastic) phase.

Materials with Unusual Behaviors or Unusual Materials

Only a few materials fit into this category, unusually they can be categorized into one of the above material classes. Water fits into the category of materials with unusual behavior. Water is one of the few materials that expands upon freezing. It changes volume by approximately 8° 0 transiting from the liquid to the solid state. Indicating devices were produced based on the freezing and expansion of water. A glass vial was filled completely with water and nucleating agents and hermetically sealed. The vial was placed on a piece of paper impregnated on the opposite side with a water-soluble dye. Then the assembly was packaged together. The device was placed in various type of perishables, such as frozen turkeys. The water in the vial would freeze and expand until it broke the vial. Upon warming the ice would melt, wicking through the paper, wetting the dye, and causing the color to migrate to the other side. The appearance of broken glass or the sound of broken glass when shaken was an indication that the object was at 0°C, while the appearance of a color was an indication that the object was at both 0°C and warmed to temperatures higher than 0°C. This represented an ingenious method to monitor the temperature profile of certain products. These devices are known as Freeze-Thaw Watches (40).

Fullerene and its derivatives can be included in the usual material category. Fullerenes are a spherically caged molecules with carbon atoms at the corner of a polyhedral structure consisting of pentagons and hexagons, the stable of which is known as buckyball or buckminsterfullerene, C₆₀. This family of materials has been under commercial development for several years. One interesting application of fullerene as a smart material has been in the area of embedding fullerenes into sol–gel matrices for the purpose of enhancing optical limiting properties (41). A semiconducting material with a magnetic ordering at 16.1 K was produced from the reaction of buckyball with tetra(dimethylamino)ethylene. This organic-based magnet did not have either the coercivity or the saturation magnetization (less than 4° 0 of predictions) to function totally as a ferromagnet. The reaction of tetra(dimethylamino)ethylene with higher molecular weight versions of fullerenes (carbon numbers 70, 84, 90, and 96) do not yield any complexes that showed magnetic ordering (18).

Future Considerations

The area of smart materials is an wide open field for the development of products that can take advantage of the material's ability to response to a stimuli in a timely and controlled manner. Also to a certain extent it is very confusing. To address the confusing issue first, the field may have been named improperly at first. To some when the word *smart* or *intelligent* is used, it implies that some materials are not smart. All materials are smart. How some materials respond to certain environmental stimuli may not be acceptable, but according to some definitions and concepts they are at least responding. One may apply the arguments of reversibility; however, several materials described in the open literature as smart lacked reversibility in their stimuli response actions. What is a word or are words that would better describe these materials, maybe "smarter" or "more intelligent".

What is the future of these smarter (more intelligent) materials? As one gets more involved in this smart material age, one is totally amazed by the number and types of applications that can be developed as well as the ones that have been developed. The ability of tailoring these materials to exhibit multistimuli and multiresponsive behaviors is becoming more of a key features to smart materials. In the open literature there are reports that describe the ingenious adaptations of multistimuli and multiresponsive smart materials, one of which discusses a smart gel based on the polymerization of *N*-isopropylacrylamide, a derivative of tris(2,2'-bipyridyl)ruthenium(II) that has a polymerizable vinyl group, and *N,N'*-methylenebisacrylamide, that is self-oscillating (42). It can simulate the beating a heart with color changes.

The smart materials structures field is new. As these materials are better understood, better devices can designed with them and the quality of life may be enhanced.

BIBLIOGRAPHY

1. J. T. Buckley, *USA Today*, *Section A*, 1–2 (Jan. 16, 1997).

2. I. Amato, *Science News*, **137**(10), 152–155 (March 10, 1990).

3. O. Port, *Business Week* (3224), 48–55 (July 29, 1991).

4. Committee on New Sensor Technologies: Materials and Applications, National Materials Advisory Board, Commission on Engineering and Technical Systems, *National Research Council Report: Expanding the Vision of Sensor Materials*, National Academy Press, Washington, D.C. 1995, pp. 33–45.

5. C. A. Rogers, *Scientific American* **273**(3), 122–126 (Sept. 1995).

6. I. Chopra, *Proceedings of the Society of Photo-Optical Instrumentation Engineers (Smart Structures and Materials 1996—Smart Structures and Integrated Systems)*; Proceedings of the Conference, San Diego, Calif., Feb. 26–29, 1996) **2717**, 20–26, (1996).

7. E. Udd, *Fiber Optic Smart Structures*, John Wiley & Sons, Inc., New York, 1995.

8. R. O. Claus and K. A. Murphy, *Proceedings of the Society of Photo-Optical Instrumentation Engineers 2nd International Conference on Optoelectronic Science and Engineering '94*, Beijing, China, Aug. 15–18, 1994, **2321**, 404–407 (1994).

9. J. E. Frommer and R. Chance, "Electrically Conductive Polymers" in J. I. Kroschwitz, ed., *Encyclopedia of Polymer Science and Engineering*, 2nd ed., Vol. 5, Wiley-Interscience, New York, 1988, pp. 462–507.

10. M. Stilka, "Photoconductive Polymers" in J. I. Kroschwitz, ed., *Encyclopedia of Polymer Science and Engineering*, 2nd ed., Vol. 11, Wiley-Interscience, New York, 1988, pp. 154–175.

11. J. Unsworth, C. Conn, Z. Jin, A. Kaynak, R. Ediriweera, P. Innis, and N. Booth, *J. Intelligent Mater. Systems Structures* **5**, 595–604 (Sept. 1994).

12. P. N. Prasad and D. J. Williams, *Introduction to Nonlinear Optical Effects in Molecules and Polymers*, Wiley-Interscience, New York, 1990.

13. R. W. Boyd, *Nonlinear Optics*, Academic Press, New York, 1992.

14. R. E. Newnham, *Mater. Res. Soc. Bull.* **22**,(5), 20–34 (May 1997).

15. B. Culshaw, *Smart Structures and Materials*, Artech House, Boston, 1996, pp. 43–45, 117–130.

16. S. R. White, *Proceedings of the Society of Photo-Optical Instrumentation Engineers (Smart Structures and Materials 1995-Mathematics and Control in Structures)*, Proceedings of the Conference, San Diego, Calif., Feb. 27–Mar. 1, 1995) **2442**, 337–348 (1995).

17. A. B. Frazier, M. R. Khan, and M. G. Allen, *Symposium Proceedings of the Materials Research Society (Smart Materials Fabrication and Materials for Micro-Electro-Mechanical Systems)*, San Francisco, Calif., Apr. 28–30, 1992, **276**, 295–301 (1992).

18. J. S. Miller and A. J. Epstein, *Chem. Eng. News*, 30–41 (Oct. 2, 1995).

19. R. Dagani, *Chem. Eng. News*, 30–33 (Sept. 18, 1995).

20. D. E. Bergbreiter, B. C. Ponder, G. Aguilar, and B. Srinivas, *Chem. Mater.* **9**(2), 472–477 (1997).

21. Y. G. Takei, T. Aoki, K. Sanui, N. Ogata, Y. Sakurai, and T. Okano, *Macromolecules* **27**, 6136–6136 (1994).

22. J. Jhang, R. Pelton, and Y. L. Deng, *Langmuir* **11**, 2301–2302 (1995).

23. J. L. Thomas, H. You, and D. A. Tirrell, *J. Am. Chem. Soc.* **117**, 2949–2950 (1995).

24. D. H. Carey and G. S. Ferguson, *J. Am. Chem. Soc.* **118**, 9780–9781 (1996).

25. M. S. Khan, *Proceedings of the Society of Photo-Optical Instrumentation Engineers (Metal/ Nonmetal Microsystems: Physics, Technology, and Applications)*; Proceedings of the Workshop, Polanica Zdroj, Poland, Sept. 11–14, 1995, **2780**, 56–59 (1996).

26. A. Vannikov and A. Karasev, *Proceedings of the Society of Photo-Optical Instrumentation Engineers (Smart Structures and Materials 1996—Smart Electronics and MEMS)*, Proceedings Conference, San Diego, Calif., Feb. 28–29, 1996, **2722**, 252–255 (1996).

27. I. Y. Galaev, *Russian Chem. Rev.* **65**(5), 471–489 (1995).

28. A. S. Hoffman, *Artificial Organs* **19**(5), 458–467 (1995).

29. A. L. Robinson, *Science*, **200**, 1371–1374 (June 23, 1978).

30. A. J. Lovinger, *Science*, **220**, 1115–1121 (June 10, 1983).

31. G. Washington, *Smart Mater. Structures* **5**(6), 801–805 (Dec. 1996).

32. R. Dagani, *Chem. Eng. News*, 26–37 (June 9, 1997).

33. R. S. Hadland, R. K. Prud'homme, *Polyelectrolyte Gels, Properties and Applications*, American Chemical Society, Washington, D.C., 1992.

34. D. DeRossi, K. Kajiware, Y. Osada, and A. Yamauchi, *Polymer Gels Fundamental and Biomedical Applications*, Plenum, New York, 1991.

35. A. Dupommie, L. Merle-Aubry, and Y. Merle, and E. Signy, *Makromol. Chem.* **187**, 211 (1986).

36. R. Baum, *Chem. Eng. News*, 7 (Jan. 30, 1996).

37. K. Ullakko, P. G. Yakovenko, and V. G. Gavriljuk, *Proceedings of the Society of Photo-Optical Instrumentation Engineers (Smart Structures and Materials 1996—Mathematics and Control in Structures)*, Proceedings of the Conference, San Diego, Calif., Feb. 26–29, 1996 **2715**, 42–50 (1996).

38. G. Kauffman and I. Mayo, *Chem. Matters*, 4–7 (Oct. 1993).

39. D. Stoeckel and W. Yu, *Superelastic Nickel-Titanium Wire*, Technical Brochure, Raychem Corporation, Menlo Park, Calif.

40. **U.S. Pat. 4,145,918**, T. W. Couch, E. P. Fournier, J. A. Harvey, D. T. A. Hibers, and H. A. Mercer.

41. R. Signorini, M. Zerbetto, M. Meneghetti, R. Bozio, G. Brusatin, E. Menegazzo, M. Guglielmi, M. Maggini, G. Scorrano, and M. Prato, *Proceedings of the Society of Photo-Optical Instrumental Engineers, Fullerenes and Photoinics III*, **2854**, 130–139 (1996).

42. R. Yoshida, T. Takahashi, T. Yamaguchi, and H. Ichijo, *Advanced Mater.* **9**, 175 (1997).

James A. Harvey
Hewlett-Packard Company and Oregon Graduate Institute of Science & Technology

SOLVENT RECOVERY, CONDENSATION

As had been noted in the previous edition of this *Encyclopedia*, the removal of solvents by condensing them from an inert gas in a water-cooled condenser is inefficient, and may create a potentially hazardous condition at the point of discharge. Vapor pressures of most solvents are relatively high. Therefore, at temperatures obtainable in a water-cooled heat exchanger, these substances maintain an appreciable partial pressure in the discharging gas. An environmentally unacceptable high VOC (volatile organic compound) level in the emitted gas, and the fact that the remaining solvent concentration is in many cases above the lower explosive limit, makes a water-cooled heat exchanger an unlikely instrument for solvent removal to an environmentally acceptable level.

Chilling the vent stream with a refrigerated solution can appreciably reduce the discharge of VOCs, but requires detailed calculations to assure that any solvents remaining in the exiting gas stream is within acceptable limits. Reducing the gas temperature to 25°C (3°F) may be economically viable if it accomplishes this goal. The following examples illustrate the reductions that are feasible when mechanical refrigeration is applied.

Solvent mole fraction	Equilibrium mole fraction in air at 25°C	Equilibrium mole fraction in air at 25°C
methanol	0.075	0.0065
benzene	0.13	0.004
acetone	0.30	0.019

If regulations governing specific emission limit VOC concentrations to the low ppm range then, of course, vapor fractions such as those illustrated by the above tabulation will not be acceptable. It may, however, still be justified to consider VOC condensation as a precursor to a final abatement device such as an adsorption bed. Removing most of the solvent from a vent stream by condensation, can drastically reduce the size and cost of a downstream cleanup system.

For optimum solvent removal, the heat exchanger should be arranged so the exiting gas temperature approaches that of the incoming cooling medium. This is best accomplished in a reflux condenser (dephlegmator) wherein solvent laden vapor is introduced into the bottom head of a vertical tubular heat exchanger, and lean vapor leaves from the top. The condensed solvent will flow countercurrent to an upflowing gas and is withdrawn from the bottom head. Cooling medium is introduced at the top of the shell to assure that uncondensed gas leaves at the lowest possible temperature.

The principal concern in designing a tubeside reflux condenser is that an excessive velocity of the entering gas will prevent downflow of the condensed solvent. This is usually not a problem in atmospheric condensation. However, because of the higher gas velocities at low pressures, it may cause flooding under vacuum conditions unless an adequate number of tubes is provided to lower the velocity. This can produce a heat-exchanger configuration that uses many more tubes (albeit shorter ones) than might ordinarily be required were the design based strictly on heat-transfer considerations. A horizontal shell side reflux condenser is less prone to flooding, but is usually not as effective in lowering the exiting gas temperature. A demonstrated procedure for designing in-tube reflux condensers that will not flood was developed by Holmes and has been published in *Perry's Chemical Engineers' Handbook* (1).

It is the author's judgment that the thermal design of a vent condenser is best accomplished by estimating an overall heat-transfer coefficient based on past experience in a similar service. Although detailed design procedures for sizing partial condensers are available through such organizations as HTRI, HTFS, B-Jac, etc, it is very difficult to determine accurately the actual amounts of inerts and VOCs that are being discharged from a process. Employing a complex calculation procedure while using imprecise input data cannot be justified.

If solvent recovery is maximized by minimizing the temperature approach, the overall heat-transfer coefficient in the condenser will be reduced. This is due to the fact that a large fraction of the heat transfer area is now utilized for cooling a gas rather than condensing a liquid. Depending on the desired temperature approach the overall heat-transfer coefficients in vent condensers usually range between 85 and 170 W /m²K (ca 15 and 30 Btu /h·ft.²°F).

BIBLIOGRAPHY

"Solvent Recovery" in *ECT* 1st ed., Vol. 12, pp. 641–654, by C. M. Cooper, Michigan State College; "Solvent Recovery" in *ECT* 2nd ed., Vol. 18, pp. 549–564, by C. M. Cooper, Michigan State University; "Solvent Recovery" in *ECT* 3rd ed., by C. M. Cooper, Michigan State University.

1. *Perry's Chemical Engineers Handbook*, 6th ed., McGraw-Hill, New York, 1984, pp. 5–59.

Otto Frank
Air Products and Chemicals, Inc.

SONOCHEMISTRY

Ultrasonic irradiation of liquids causes high energy chemical reactions to occur, often with the emission of light (1–5). The origin of sonochemistry and sonoluminescence is acoustic cavitation: the formation, growth, and implosive collapse of bubbles in liquids irradiated with high intensity sound. The collapse of bubbles caused by cavitation produces intense local heating and high pressures, with very short lifetimes. In clouds of cavitating bubbles, these hot-spots (6,7) have equivalent temperatures of roughly 5000 K, pressures of about 1000 atmospheres, and heating and cooling rates above 10^{10} K s. In single-bubble cavitation, conditions may be even more extreme. Thus, cavitation can create extraordinary physical and chemical conditions in otherwise cold liquids.

When liquids that contains solids are irradiated with ultrasound, related phenomena can occur. When cavitation occurs near an extended solid surface, cavity collapse is nonspherical and drives high speed jets of liquid into the surface (8,9). These jets and associated shock waves can cause substantial surface damage and expose fresh, highly heated surfaces. Ultrasonic irradiation of liquid–powder suspensions produces another effect: high velocity interparticle collisions. Cavitation and the shockwaves it creates in a slurry can accelerate solid particles to high velocities (10). The resultant collisions are capable of inducing dramatic changes in surface morphology, composition, and reactivity (11).

Sonochemistry can be roughly divided into categories based on the nature of the cavitation event: homogeneous sonochemistry of liquids, heterogeneous sonochemistry of liquid–liquid or liquid–solid systems, and sonocatalysis (which overlaps the first two) (12–15). In some cases, ultrasonic irradiation can increase reactivity by nearly a million-fold (16). Because cavitation can only occur in liquids, chemical reactions are not generally seen in the ultrasonic irradiation of solids or solid-gas systems.

Sonoluminescence in general may be considered a special case of homogeneous sonochemistry; however, recent discoveries in this field have heightened interest in the phenomenon in and by itself (17,18). Under conditions where an isolated, single bubble undergoes cavitation, recent studies on the duration of the sonoluminescence flash suggest that a shock wave may be created within the collapsing bubble, with the capacity to generate truly enormous temperatures and pressures within the gas.

Acoustic Cavitation

The chemical effects of ultrasound do not arise from a direct interaction with molecular species. Ultrasound spans the frequencies of roughly 15 kHz to 1 GHz. With sound velocities in liquids typically about 1500 m s, acoustic wavelengths range from roughly 10 to 10^{-4} cm. These are not molecular dimensions. Consequently, no direct coupling of the acoustic field with chemical species on a molecular level can account for sonochemistry or sonoluminescence.

Instead, sonochemistry and sonoluminescence derive principally from acoustic cavitation (9), which serves as an effective means of concentrating the diffuse energy of sound. Compression of a gas generates heat. The compression of bubbles during cavitation is more rapid than thermal transport, which generates a short-lived, localized hot-spot. There is a general consensus that this hot-spot is the source of homogeneous sonochemistry. Rayleigh's early description of a mathematical model for the collapse of cavities in incompressible liquids predicted enormous local temperatures and pressures (19). Ten years later, Richards and Loomis reported the first chemical and biological effects of ultrasound (20). Alternative mechanisms involving electrical microdischarge have been occasionally proposed (21,22), but remain a minority viewpoint.

If the acoustic pressure amplitude of a propagating acoustic wave is relatively large (greater than ≈ 0.5 MPa), local inhomogeneities in the liquid (eg, gas-filled crevices in particulates) can give rise to the explosive growth of a nucleation site into a cavity of macroscopic dimensions, primarily filled with vapor. Such a bubble is inherently unstable, and its subsequent collapse can result in an enormous concentration of energy (Fig. 1). This violent cavitation event has been termed transient cavitation (23). A normal consequence of this unstable growth and subsequent collapse is that the cavitation bubble itself is destroyed. Gas-filled remnants from the collapse, however, may give rise to reinitiation of the process.

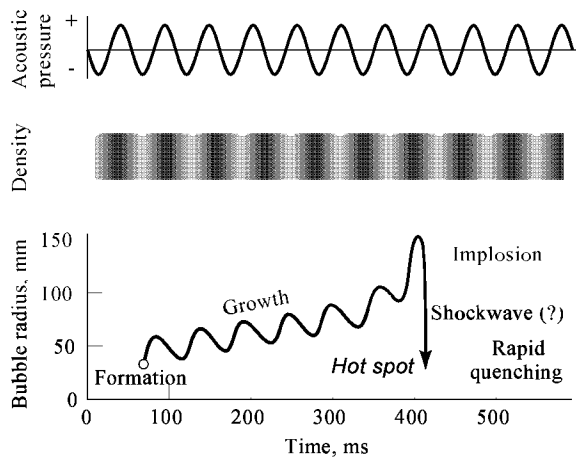


Fig. 1. Transient acoustic cavitation: the origin of sonochemistry and sonoluminescence.

The generally accepted explanation for the origin for the origin of sonochemistry and sonoluminescence is the hot-spot theory, in which the potential energy given the bubble as it expands to maximum size is concentrated into a heated gas core as the bubble implodes. The oscillations of a gas bubble driven by an acoustic field are generally described by Rayleigh-Plesset equation; one form of which, called the Gilmore equation (9,23), can be expressed a second-order nonlinear differential equation given as

$$R \left(1 + \frac{U}{C} \right) \frac{d^2 R}{dt^2} + \frac{3}{2} \left(1 + \frac{U}{3C} \right) \left(\frac{dR}{dt} \right)^2 - \left(1 + \frac{U}{C} \right) H - \frac{R}{C} \left(1 + \frac{U}{C} \right) \frac{dH}{dt} = 0 \quad (1)$$

The radius and velocity of the bubble wall are given by R and U respectively. The values for H , the enthalpy at the bubble wall, and C , the local sound speed, may be expressed as follows, using the Tait equation of state for the liquid.

$$H = \frac{n}{n-1} \frac{A^{1-n}}{\rho_o} \left[(P(R) + B)^{n-1-n} - (P_\infty(t) + B)^{n-1-n} \right]$$

(2)

and

$$C = \left[c_o^2 + (n-1)H \right]$$

(3)

The linear speed of sound in the liquid is c_o . A , B , and n are constants that should be set to the appropriate values for water. Any acoustic forcing function is included in the pressure at infinity term, $P_\infty(t)$. The pressure at the bubble wall, $P(R)$, is given by

$$P(R) = \left(P_o + \frac{2\sigma}{R} \right) \left(\frac{R_o}{R} \right)^{3\gamma} - \frac{2\sigma}{R} - \frac{4\mu U}{R}$$

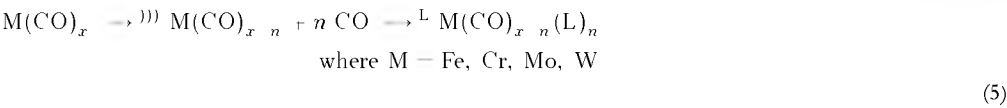
(4)

where the initial radius of the bubble at time zero is R_o . The ambient pressure of the liquid is P_o , the surface tension σ , the shear viscosity μ , and the polytropic exponent γ .

The validity of the Gilmore equation to compute the behavior of a single, isolated cavitating bubble has been experimentally confirmed. For example, using a light-scattering technique, various researchers have obtained measurements of the radius–time curve for single collapsing bubbles, simultaneous with optical emission from sonoluminescence (see below). The single-bubble sonoluminescent emission is seen as a sharp spike, appearing at the final stages of bubble collapse, and the general shape of the theoretical radius-time curve is observed (24–26).

Two-Site Model of Sonochemical Reactivity

The transient nature of the cavitation event precludes conventional measurement of the conditions generated during bubble collapse. Chemical reactions themselves, however, can be used to probe reaction conditions. The effective temperature realized by the collapse of clouds of cavitating bubbles can be determined by the use of competing unimolecular reactions whose rate dependencies on temperature have already been measured. This technique of comparative-rate chemical thermometry was used by Suslick, Hammerton, and Cline to first determine the effective temperature reached during cavity collapse (6). The sonochemical ligand substitutions of volatile metal carbonyls were used as these comparative rate probes (eq. 5, where the symbol $\rightarrow$ represents ultrasonic irradiation of a solution, and L represents a substituting ligand). These kinetic studies revealed that there were in fact



two sonochemical reaction sites: the first (and dominant site) is the bubble's interior gas-phase while the second is an *initially* liquid phase. The latter corresponds either to heating of a shell of liquid around the collapsing bubble or to droplets of liquid ejected into the hot-spot by surface wave distortions of the collapsing bubble, as shown schematically in Figure 2.

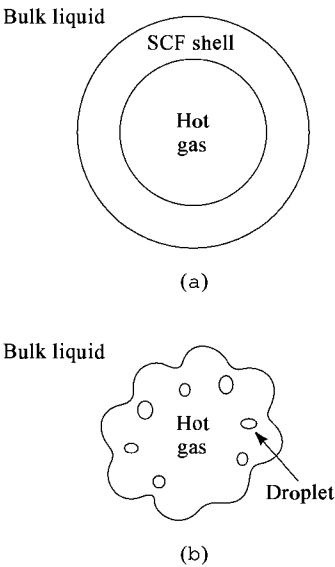


Fig. 2. Two-site models of the sonochemical reactions sites. (a) Thermal diffusion shell model; (b) surface wave droplet model.

The effective local temperatures in both sites were determined. By combining the relative sonochemical reaction rates for equation 5 with the known temperature behavior of these reactions, the conditions present during cavity collapse could then be calculated. The effective temperature of these hotspots was measured at ≈ 5200 K in the gas-phase reaction zone and ≈ 1900 K in the initially liquid zone (6). Of course, the comparative rate data represent only a composite temperature: during the collapse, the temperature has a highly dynamic profile, as well as a spatial temperature gradient. This two-site model has been confirmed with other reactions (27,28) and alternative measurements of local temperatures by sonoluminescence are consistent (7), as discussed later.

Microjet Formation during Cavitation at Liquid–Solid Interfaces

A very different phenomenon arises when cavitation occurs near extended liquid–solid interfaces. There are two proposed mechanisms for the effects of cavitation near surfaces: microjet impact and shockwave damage. Whenever a cavitation bubble is produced near a boundary, the asymmetry of the liquid particle motion during cavity collapse can induce a strong deformation in the cavity (9). The potential energy of the expanded bubble is converted into kinetic energy of a liquid jet that extends through the bubble's interior and penetrates the opposite bubble wall. Because most of the available energy is transferred to the accelerating jet, rather than the bubble wall itself, this jet can reach velocities of hundreds of meters per second. Because of the induced asymmetry, the jet often impacts the solid boundary and can deposit enormous energy densities at the site of impact. Such energy concentration can result in severe damage to the boundary surface. Figure 3 is a photograph of a jet developed in a collapsing cavity. The second mechanism of cavitation-induced surface damage invokes shockwaves created by cavity collapse in the liquid. The impingement of microjets and shockwaves on the surface creates the localized erosion responsible for much of ultrasonic cleaning and many of the sonochemical effects on heterogeneous reactions. In this process, the erosion of metals by cavitation generates newly exposed, highly heated surfaces that are highly reactive.



Fig. 3. Liquid jet produced during collapse of a cavitation bubble near a solid surface. The width of the bubble is about 1 mm.

Reproduced with permission (8).

A solid surface several times larger than the resonance bubble size is necessary to induce distortions during bubble collapse. For ultrasound of ≈ 20 kHz, damage associated with jet formation cannot occur if the solid particles are smaller than $\approx 200\text{ }\mu\text{m}$. In these cases, however, the shockwaves created by homogeneous cavitation can create high velocity interparticle collisions (10,11). Suslick and co-workers have found that the turbulent flow and shockwaves produced by intense ultrasound can drive metal particles together at sufficiently high speeds to induce effective melting in direct collisions (Fig. 4) and the abrasion of surface crystallites in glancing impacts (Fig. 5). A series of transition-metal powders were used to probe the maximum temperatures and speeds reached during interparticle collisions. Using the irradiation of Cr, Mo, and W powders in decane at 20 kHz and 50 W cm^2 , agglomeration and essentially a localized melting occurs for the first two metals, but not the third (Fig. 6). On the basis of the melting points of these metals, the effective transient temperature reached at the point of impact during interparticle collisions is roughly 3000°C (which is unrelated to the temperature inside the hot-spot of a collapsing bubble). From the volume of the melted region of impact, the amount of energy generated during collision was determined. From this, a lower estimate of the velocity of impact is roughly one half the speed of sound (10). These are precisely the effects expected on suspended particulates from cavitation-induced shockwaves in the liquid.

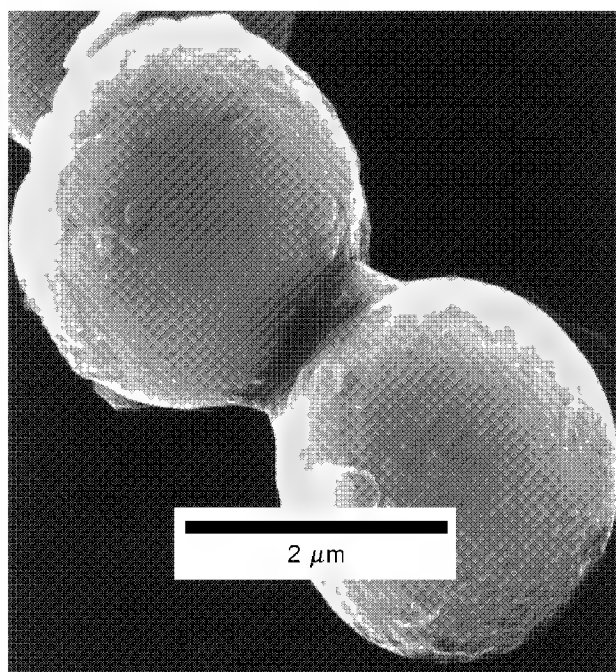
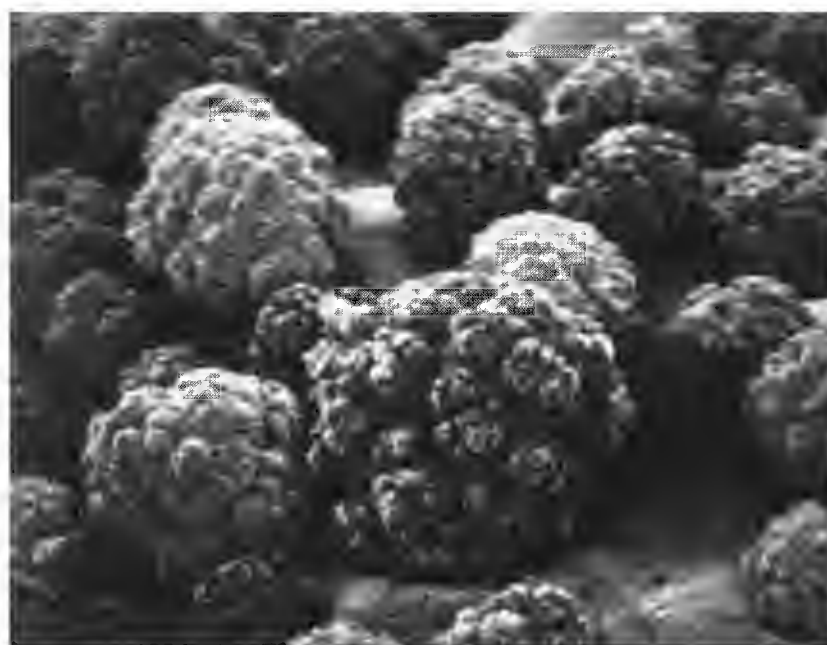
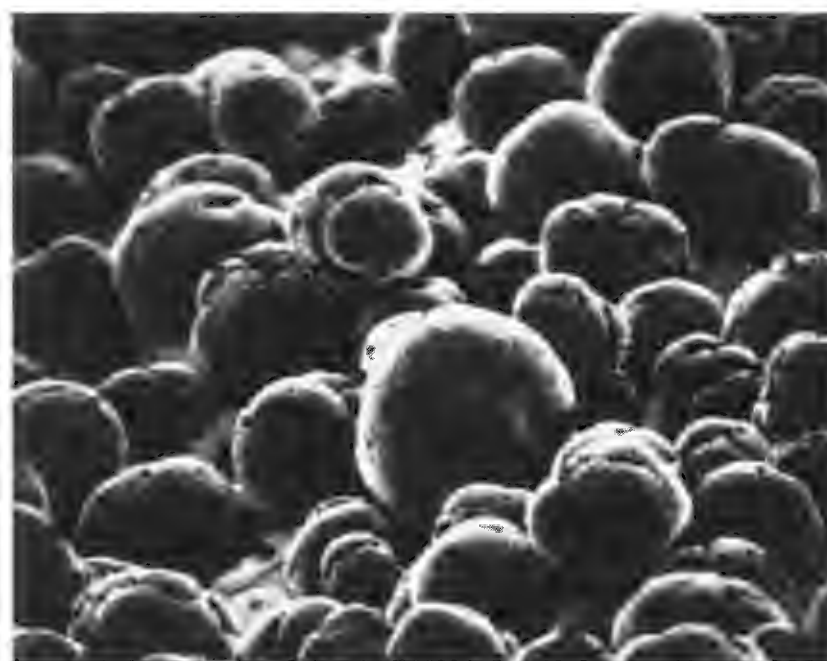


Fig. 4. Scanning electron micrograph of 5- μm diameter Zn powder. Neck formation from localized melting is caused by high-velocity interparticle collisions. Similar micrographs and elemental composition maps (by Auger electron spectroscopy) of mixed metal collisions have also been made.

Reproduced with permission (10).



(a)



(b)

Fig. 5. The effect of ultrasonic irradiation on the surface morphology and particle size of Ni powder. Initial particle diameters (a) before ultrasound were $\approx 160\text{ }\mu\text{m}$; (b) after ultrasound, $\approx 80\text{ }\mu\text{m}$. High velocity interparticle collisions caused by ultrasonic irradiation of slurries are responsible for the smoothing and removal of passivating oxide coating.

Reproduced with permission (11).

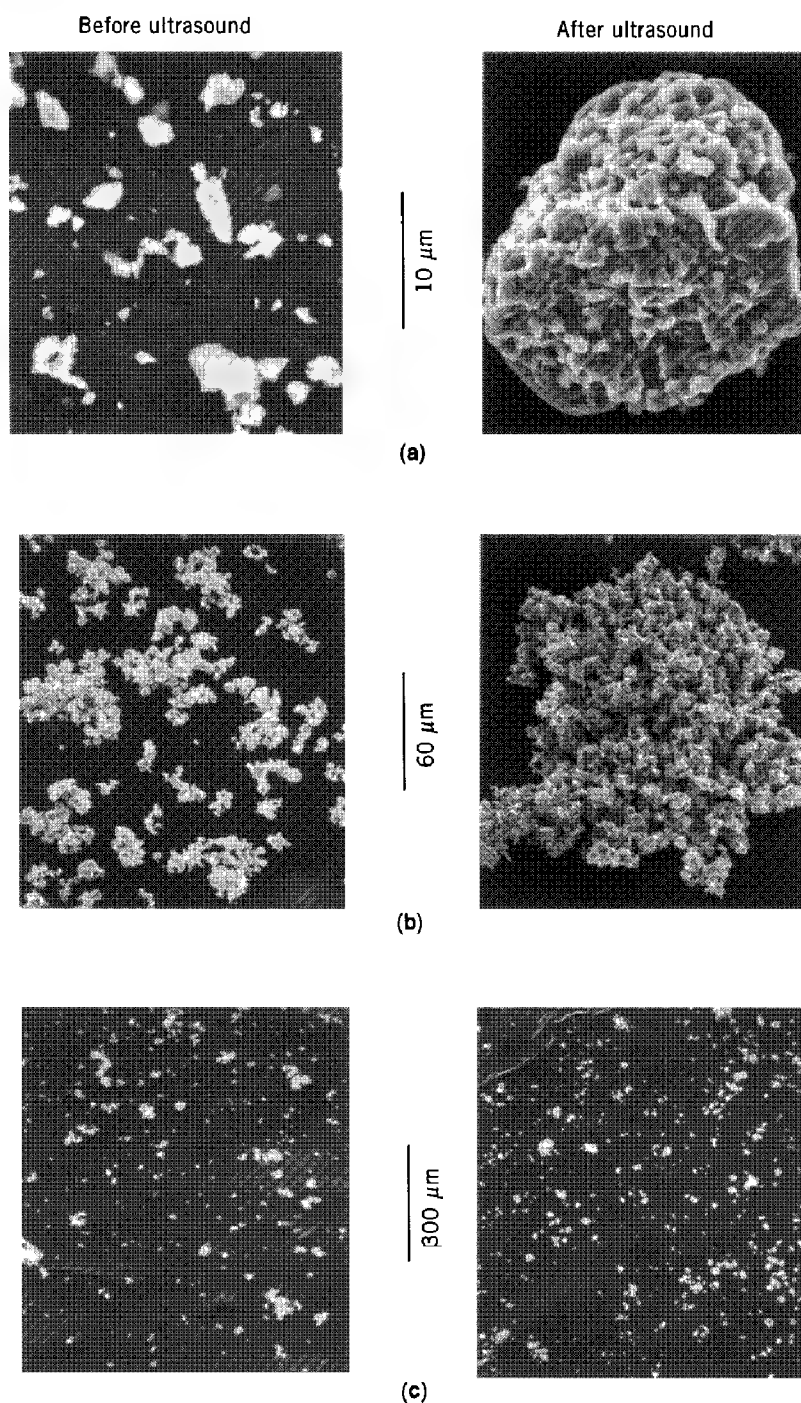


Fig. 6. The effect of ultrasound on particle agglomeration of (a) Cr (mp 1857°C), (b) Mo (mp 2617°C), and (c) W (mp 3410°C) after irradiation of decane slurries under Ar.

Reproduced with permission (10).

Sonoluminescence

Types of Sonoluminescence. In addition to driving chemical reactions, ultrasonic irradiation of liquids can also produce light. Sonoluminescence was first observed from water in 1934 by Frenzel and Schultes (29). As with sonochemistry, sonoluminescence derives from acoustic cavitation. It is now generally thought that there are two separate forms of sonoluminescence: multiple-bubble sonoluminescence (MBSL) and single-bubble sonoluminescence (SBSL) (17,24,30). Since cavitation is a nucleated process and liquids generally contain large numbers of particulates that serve as nuclei, the cavitation field generated by propagating or standing acoustic wave typically consists of very large numbers of interacting bubbles, distributed over an extended region of the liquid. If this cavitation is sufficiently intense to produce sonoluminescence, then this phenomenon is called multiple-bubble sonoluminescence (MBSL) (2,17).

Under the appropriate conditions, the acoustic force on a bubble can be used to balance against its buoyancy, holding the single bubble isolated in the liquid by acoustic levitation. This permits examination of the dynamic characteristics of the bubble in considerable detail, from both a theoretical and an experimental perspective. Such a bubble is typically quite small, compared to an acoustic wavelength (eg, at 20 kHz, the resonance size is approximately 150 μm). It was recently discovered for rather specialized but easily obtainable conditions, that a single, stable, oscillating gas bubble can be forced into such large amplitude pulsations that it produces sonoluminescence emissions on each (and every) acoustic cycle (31,32). This phenomenon is called single-bubble sonoluminescence, and has received considerable recent attention (17,18,33,34).

Multiple-Bubble Sonoluminescence. The sonoluminescence of aqueous solutions has been often examined over the past thirty years. The spectrum of MBSL in water consists of a peak at 310 nm and a broad continuum throughout the visible region. An intensive study of aqueous MBSL was conducted by Verrall and Sehgal (35). The emission at 310 nm is from excited-state $\text{OH}^\bullet$, but the continuum is difficult to interpret. MBSL from aqueous and alcohol solutions of many metal salts have been reported and are characterized by emission from metal atom excited states (36).

Sonoluminescence from nonaqueous liquids has only recently been examined. Flint and Suslick reported the first MBSL spectra of organic liquids (37). With various hydrocarbons, the observed emission is from excited states of C_2 ($d^3\Pi_g \rightarrow a^3\Pi_u$, the Swan lines), the same emission seen in flames. Furthermore, the ultrasonic irradiation of alkanes in the presence of N_2 (or NH_3 or amines) gives emission from CN excited states, but not from N_2 excited states. Emission from N_2 excited states would have been expected if the MBSL originated from microdischarge, whereas CN emission is typically observed from thermal sources. When oxygen is present, emission from excited states of CO_2 , $\text{CH}^\bullet$, and $\text{OH}^\bullet$ is observed, again similar to flame emission.

For both aqueous and nonaqueous liquids, MBSL is caused by chemical reactions of high energy species formed during cavitation by bubble collapse, and its principal source is most probably not blackbody radiation or electrical discharge. MBSL is predominantly a form of chemiluminescence.

Single-Bubble Sonoluminescence. The spectra of MBSL and SBSL are dramatically different. MBSL is generally dominated by atomic and molecular emission lines, but SBSL is an essentially featureless emission that increases with decreasing wavelength. For example, an aqueous solution of NaCl shows evidence of excited states of both OH· and Na in the MBSL spectrum; however, the SBSL spectrum of an identical solution shows no evidence of either of these peaks (30). Similarly, the MBSL spectrum falls off at low wavelengths, while the SBSL spectrum continues to rise, at least for bubbles containing most noble gases (38).

An intriguing aspect of SBSL is the extremely short duration of the sonoluminescence flash. The hydrodynamic models of adiabatic collapse of a single bubble suggest that the temperature of the gas within the bubble should remain at elevated temperatures for times on the order of tens of nanoseconds; however, there is strong evidence that the pulse duration of the SBSL flash is three orders of magnitude shorter. Putterman and his colleagues, using the fastest PMT available, reported that this duration is less than 50 ps, perhaps much less (39). The most plausible explanation for this short flash interval, and some of the observed spectra (see below), is that an imploding shock wave is created within the gas bubble during the final stages of collapse. If this shock wave does indeed exist, exciting possibilities can be inferred about the temperatures that could be attained within the bubble and the physics that might result. Indeed, speculations on the possibilities of inertial confinement (*hot*) fusion have been made (40,41).

Spectroscopic Probes of Cavitation Conditions. Determination of the temperatures reached in a cavitating bubble has remained a difficult experimental problem. As a spectroscopic probe of the cavitation event, MBSL provides a solution. High resolution MBSL spectra from silicone oil under Ar have been reported and analyzed (7). The observed emission comes from excited state C_2 and has been modeled with synthetic spectra as a function of rotational and vibrational temperatures, as shown in Figure 7. From comparison of synthetic to observed spectra, the effective cavitation temperature is 5050 ± 150 K. The excellence of the match between the observed MBSL and the synthetic spectra provides definitive proof that the sonoluminescence event is a thermal, chemiluminescence process. The agreement between this spectroscopic determination of the cavitation temperature and that made by comparative rate thermometry of sonochemical reactions is surprisingly close (6).

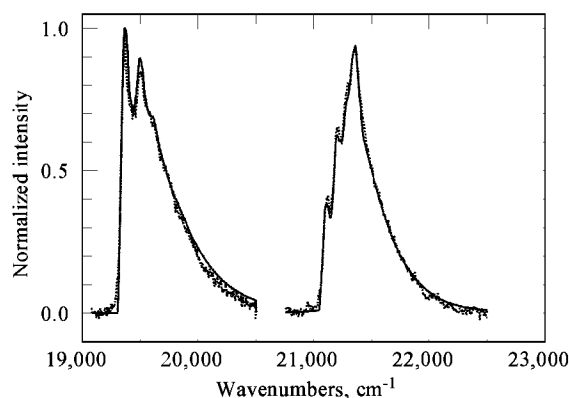


Fig. 7. Sonoluminescence of excited state C_2 . Emission from the $\Delta v = +1$ manifold of the $d^3\Pi_g - a^3\Pi_u$ transition (Svan band) of C_2 . Reproduced with permission (7). ·····, Observed sonoluminescence from polydimethylsiloxane silicone oil under Ar at 0°C; —, best fit synthetic spectrum, with $T_v = T_r = 4900$ K.

The interpretation of the spectroscopy of SBSL is much less clear. At this writing, SBSL has been observed primarily in aqueous fluids, and the spectra obtained are surprisingly featureless. Some very interesting effects are observed when the gas contents of the bubble are changed (39,42). Furthermore, the spectra show practically no evidence of OH emissions, and when He and Ar bubbles are considered, continue to increase in intensity even into the deep ultraviolet. These spectra are reminiscent of blackbody emission with temperatures *considerably* in excess of 5000 K and lend some support to the concept of an imploding shock wave (41). Several other alternative explanations for SBSL have been presented, and there exists considerable theoretical activity in this particular aspect of SBSL.

Sonochemistry

In a fundamental sense, chemistry is the interaction of energy and matter. Chemical reactions require energy in one form or another to proceed: chemistry stops as the temperature approaches absolute zero. One has only limited control, however, over the nature of this interaction. In large part, the properties of a specific energy source determines the course of a chemical reaction. Ultrasonic irradiation differs from traditional energy sources (such as heat, light, or ionizing radiation) in duration, pressure, and energy per molecule. The immense local temperatures and pressures and the extraordinary heating and cooling rates generated by cavitation bubble collapse mean that ultrasound provides an unusual mechanism for generating high energy chemistry. Like photochemistry, very large amounts of energy are introduced in a short period of time, but it is thermal, not electronic, excitation. As in flash pyrolysis, high thermal temperatures are reached, but the duration is very much shorter (by $>10^4$) and the temperatures are even higher (by five- to tenfold). Similar to shock-tube chemistry or multiphoton infrared laser photolysis, cavitation heating is very short lived, but occurs within condensed phases. Furthermore, sonochemistry has a high-pressure component, which suggests that one might be able to produce on a microscopic scale the same macroscopic conditions of high temperature–pressure “bomb” reactions or explosive shockwave synthesis in solids. Figure 8 presents an interesting comparison of the parameters that control chemical reactivity (time, pressure, and energy) for various forms of chemistry.

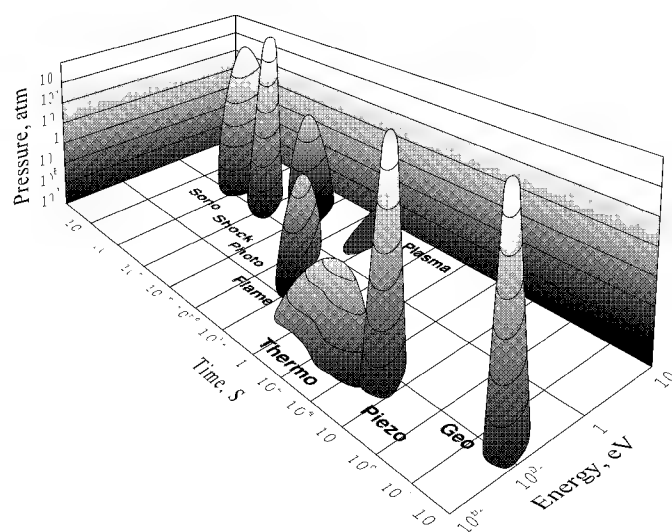


Fig. 8. Chemistry: the interaction of energy and matter. To convert atm to Pa, multiply by 1.013×10^5 .

Experimental Design. A variety of devices have been used for ultrasonic irradiation of solutions. There are three general designs in use presently: the ultrasonic cleaning bath, the direct immersion ultrasonic horn, and flow reactors. The originating source of the ultrasound is generally a piezoelectric material, usually a lead zirconate titanate ceramic (PZT), which is subjected to a high ac voltage with an ultrasonic frequency (typically 15 to 50 kHz). For industrial use, the more robust magnetostrictive metal alloys (usually of Ni) can be used as the core of a solenoid generating an alternating magnetic field with an ultrasonic frequency. The vibrating source is attached to the wall of a cleaning bath, to an amplifying horn, or to the outer surfaces of a flow-through tube or diaphragm.

The ultrasonic cleaning bath is clearly the most accessible source of laboratory ultrasound and has been used successfully for a variety of liquid–solid heterogeneous sonochemical studies. Lower acoustic intensities can often be used in liquid–solid heterogeneous systems, because of the reduced liquid tensile strength at the liquid–solid interface. For such reactions, a common ultrasonic cleaning bath will therefore often suffice. The low intensity available in these devices ($\approx 1 \text{ W cm}^{-2}$), however, can prove limiting. In addition, the standing wave patterns in ultrasonic cleaners require accurate positioning of the reaction vessel. On the other hand, ultrasonic cleaning baths are easily accessible, relatively inexpensive, and usable on moderately large scale. Even in the case of heterogeneous sonochemistry, however, the ultrasonic cleaning bath must be viewed as an apparatus of limited capability.

The most intense and reliable source of ultrasound generally used in the chemical laboratory is the direct immersion ultrasonic horn (50 to 500 W cm^{-2}), as shown in Figure 9, which can be used for work under either inert or reactive atmospheres or at moderate pressures (<10 atmospheres). These devices are available from several manufacturers at modest cost and are often used by biochemists for cell disruption. A variety of sizes of power supplies and titanium horns are available, thus allowing flexibility in sample size. Commercially available flow-through reaction chambers which will attach to these horns allow the processing of multiliter volumes. The acoustic intensities are easily and reproducibly variable; the acoustic frequency is well controlled, albeit fixed (typically at 20 kHz). Since power levels are quite high, counter-cooling of the reaction solution is essential to provide temperature control. Erosion of the titanium tip is a potential disadvantage, especially in corrosive media. Such erosion is generally a very slow process without chemical consequences (given the high tensile strength and low reactivity of Ti metal) and can be avoided by using the horn to irradiate through a cooling solution into a reaction solution held in a glass container (a so-called *cup-horn*).

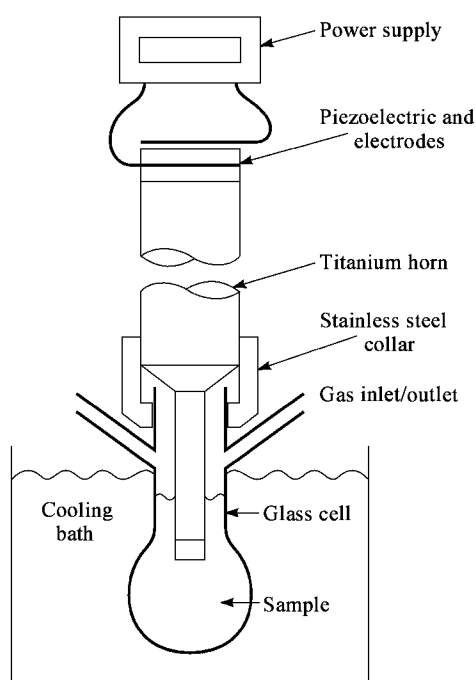


Fig. 9. A typical sonochemical apparatus with direct immersion ultrasonic horn. Ultrasound can be easily introduced into a chemical reaction with good control of temperature and ambient atmosphere. The usual piezoelectric ceramic is PZT, a lead zirconate titanate ceramic. Similar designs for sealed stainless steel cells can operate at pressures above 10 bar.

Large-scale ultrasonic generation in flow-through configurations is a well-established technology (43–46). Liquid processing rates as high as 200 L min are routinely accessible from a variety of modular, in-line designs with acoustic power of $\approx 20 \text{ kW}$ per unit. The industrial uses of these units include (1) degassing of liquids, (2) dispersion of solids into liquids, (3) emulsification of immiscible liquids, and (4) large-scale cell disruption (45,46).

Homogeneous sonochemistry typically is not a very energy efficient process (although it can be more efficient than photochemistry), whereas heterogeneous sonochemistry is several orders of magnitude better. Unlike photochemistry, whose energy inefficiency is inherent in the production of photons, ultrasound can be produced with nearly perfect efficiency from electric power. A primary limitation of sonochemistry remains the small fraction

of the acoustic power actually involved in the cavitation events. This might be significantly improved, however, if a more efficient means of coupling the sound field to generate cavitation can be found.

Sonochemistry is strongly affected by a variety of external variables, including acoustic frequency, acoustic intensity, bulk temperature, static pressure, ambient gas, and solvent (47). These are the important parameters which need consideration in the effective application of ultrasound to chemical reactions. The origin of these influences is easily understood in terms of the hot-spot mechanism of sonochemistry.

The frequency of the sound field is not a commonly altered variable in most sonochemistry. Changing sonic frequency alters the resonant size of the cavitation event and to some extent, the lifetime of the bubble collapse, but the overall process remains unchanged. Sonochemistry is therefore less influenced over the range where cavitation can occur (from tens of Hz to a few MHz). The observed sonochemical rates may change, but well-controlled comparisons of efficiency are lacking at this time and will prove difficult. Subtle differences in product distributions from homogeneous reactions have been occasionally reported (48). At very high frequencies (above a few MHz), cavitation ceases, and sonochemistry is generally not observed.

Acoustic intensity has a dramatic influence on the observed rates of sonochemical reactions. Below a threshold value, the amplitude of the sound field is too small to induce nucleation or bubble growth. Above the cavitation threshold, increased intensity of irradiation (from an immersion horn, for example) will increase the effective volume of the zone of liquid which will cavitate, and thus, increase the observed sonochemical rate. Furthermore, as the acoustic pressures increase, the range of bubble sizes which will undergo transient cavitation increases; this too will increase the observed sonochemical rate. It is often observed experimentally, however, that as one continues to increase acoustic amplitude, eventually rates begin to diminish. At high intensities, the cavitation of the liquid near the radiating surface becomes so intense as to produce a shroud of bubbles which will diminish the penetration of the sound into the liquid. In addition, bubble growth may become so rapid that the bubble grows beyond the size range of transient cavitation before implosive collapse can occur.

The effect of the bulk solution temperature lies primarily in its influence on the bubble content before collapse. With increasing temperature, in general, sonochemical reaction rates are *slower!* This reflects the dramatic influence which solvent vapor pressure has on the cavitation event: the greater the solvent vapor pressure found within a bubble prior to collapse, the less effective the collapse. There is generally a linear correlation of the log of the sonochemical rate and the total solvent vapor pressure (49). When secondary reactions are being monitored (as in chemical reactions occurring after initial acoustic erosion of a passivated surface), temperature will play its usual role in thermally activated chemical reactions. This explains the common observation that rates of heterogeneous sonochemistry often have an optimal reaction temperature: below this temperature, cavitation processes are improved, but secondary chemical reactions are slowed, and at higher temperatures, vice versa.

Increases in the applied static pressure increase the acoustic intensity necessary for cavitation, but if equal number of cavitation events occur, the collapse should be more intense. In contrast, as the ambient pressure is reduced, eventually the gas-filled crevices of particulate matter which serve as nucleation sites for the formation of cavitation in even "pure" liquids, will be deactivated, and therefore the observed sonochemistry will be diminished.

The choice of ambient gas will also have a major impact on sonochemical reactivity. The maximum temperature reached during cavitation is strongly dependent on the polytropic ratio ($\gamma = C_p / C_v$) of the ambient gas, which defines the amount of heat released during the adiabatic compression of that gas. Monatomic gases give much more heating than diatomic, which are much better than polyatomic gases (including solvent vapor). Sonochemical rates are also significantly influenced by the thermal conductivity of the ambient, so even the noble gases affect cavitation differently: He is generally much worse than Ar and Xe is the best; Ar is often the most cost-effective choice. In addition, sonochemical reactions will often involve the gases present in the cavitation event.

The choice of the solvent also has a profound influence on the observed sonochemistry. The effect of vapor pressure has already been mentioned. Other liquid properties, such as surface tension and viscosity, will alter the threshold of cavitation, but this is generally a minor concern. The chemical reactivity of the solvent is often much more important. No solvent is inert under the high temperature conditions of cavitation (50). One may minimize this problem, however, by using robust solvents that have low vapor pressures so as to minimize their concentration in the vapor phase of the cavitation event. Alternatively, one may wish to take advantage of such secondary reactions, for example, by using halocarbons for sonochemical halogenations. With ultrasonic irradiations in water, the observed aqueous sonochemistry is dominated by secondary reactions of $\text{OH}\cdot$ and $\text{H}\cdot$ formed from the sonolysis of water vapor in the cavitation zone (51–53).

Control of sonochemical reactions is subject to the same limitation that any thermal process has: the Boltzmann energy distribution means that the energy per individual molecule will vary widely. One does have easy control, however, over the energetics of cavitation through the parameters of acoustic intensity, temperature, ambient gas, and solvent choice. The thermal conductivity of the ambient gas (eg, a variable He–Ar atmosphere) and the overall solvent vapor pressure provide easy methods for the experimental control of the peak temperatures generated during the cavitation collapse.

Homogeneous Sonochemistry: Bond Breaking and Radical Formation. The chemical effect of ultrasound on aqueous solutions have been studied for many years. The primary products are H_2 and H_2O_2 ; there is strong evidence for various high-energy intermediates, including $\text{HO}_2\cdot$, $\text{H}\cdot$, $\text{OH}\cdot$, and perhaps e_{aq}^- . The elegant work of Riesz and collaborators used electron paramagnetic resonance with chemical spin-traps to demonstrate definitively the generation of $\text{H}\cdot$ and $\text{OH}\cdot$ during ultrasonic irradiation, even with clinical sources of ultrasound (51–53). The extensive work in Henglein's laboratory involving aqueous sonochemistry of dissolved gases has established clear analogies to combustion processes (27,28). As one would expect, the sonolysis of water, which produces both strong reductants and oxidants, is capable of causing secondary oxidation and reduction reactions, as often observed by Margulis and co-workers (54). Most recently there has been strong interest shown in the use of ultrasound for remediation of low levels of organic contamination of water (47,55,56). The $\text{OH}\cdot$ radicals produced from the sonolysis of water are able to attack essentially all organic compounds (including halocarbons, pesticides, and nitroaromatics) and through a series of reactions oxidize them fully. The desirability of sonolysis for such remediation lies in its low maintenance requirements and the low energy efficiency of alternative methods (eg, ozonolysis, uv photolysis).

In contrast, the ultrasonic irradiation of organic liquids has been less studied. Suslick and co-workers established that virtually all organic liquids will generate free radicals upon ultrasonic irradiation, as long as the total vapor pressure is low enough to allow effective bubble collapse (49). The sonolysis of simple hydrocarbons (for example, *n*-alkanes) creates the same kinds of products associated with very high temperature pyrolysis (50). Most of these products (H_2 , CH_4 , and the smaller 1-alkenes) derive from a well-understood radical chain mechanism.

The sonochemistry of solutes dissolved in organic liquids also remains largely unexplored. The sonochemistry of metal carbonyl compounds is an exception (57). Detailed studies of these systems led to important mechanistic understandings of the nature of sonochemistry. A variety of unusual reactivity patterns have been observed during ultrasonic irradiation, including multiple ligand dissociation, novel metal cluster formation, and the initiation of homogeneous catalysis at low ambient temperature (57).

Applications of Sonochemistry to Materials Synthesis. Of special interest is the recent development of sonochemistry as a synthetic tool for the creation of unusual inorganic materials (58,59). As one example, the recent discovery of a simple sonochemical synthesis of amorphous iron (Fig. 10) helped settle the longstanding controversy over its magnetic properties (60,61). More generally, ultrasound has proved extremely useful in the synthesis of a wide range of nanostructured materials, including high surface area transition metals, alloys, carbides, oxides and colloids (62,64). Sonochemical decomposition of volatile organometallic precursors in high boiling solvents produces nanostructured materials in various forms with high catalytic activities. Nanometer colloids, nanoporous high surface area aggregates, and nanostructured oxide supported catalysts can all be prepared by this general route, as shown schematically in Figure 11. For example, sonication of iron pentacarbonyl with silica generated an amorphous nanostructured Fe– SiO_2 supported catalyst. This catalyst showed higher catalytic activity for the Fischer-Tropsch synthesis compared to the conventional Fe–silica catalyst prepared by the traditional incipient wetness method. Sonochemical synthesis of high surface area alloys can be accomplished by the sonolysis of $\text{Fe}(\text{CO})_5$ and $\text{Co}(\text{CO})_2(\text{NO})$. As another example, ultrasonic irradiation of $\text{Mo}(\text{CO})_6$ produces aggregates of nanometer-sized clusters of face centered cubic molybdenum carbide. The extremely porous material had a high surface area and consisted of aggregates of $\approx 2\text{-nm}$ sized particles. The catalytic properties showed that the molybdenum carbide generated by ultrasound is an active and highly selective dehydrogenation catalyst comparable to commercial ultrafine platinum powder.

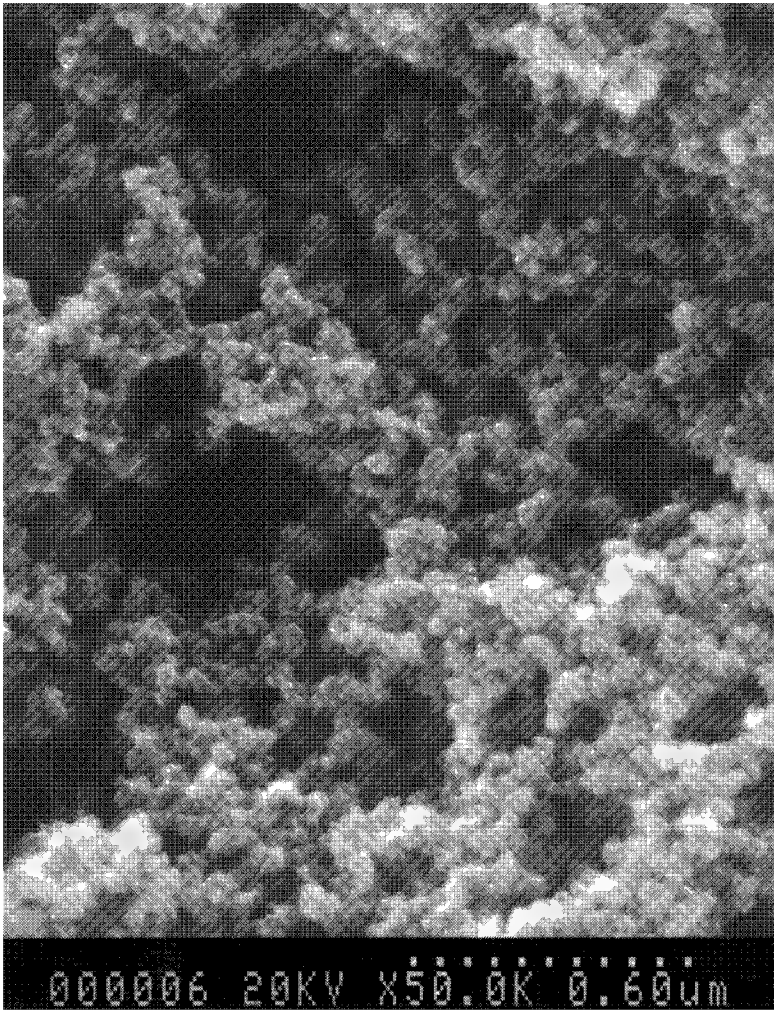


Fig. 10. Scanning electron micrograph of amorphous nanostructured iron powder produced from the ultrasonic irradiation of $\text{Fe}(\text{CO})_5$.

Reproduced with permission (60).

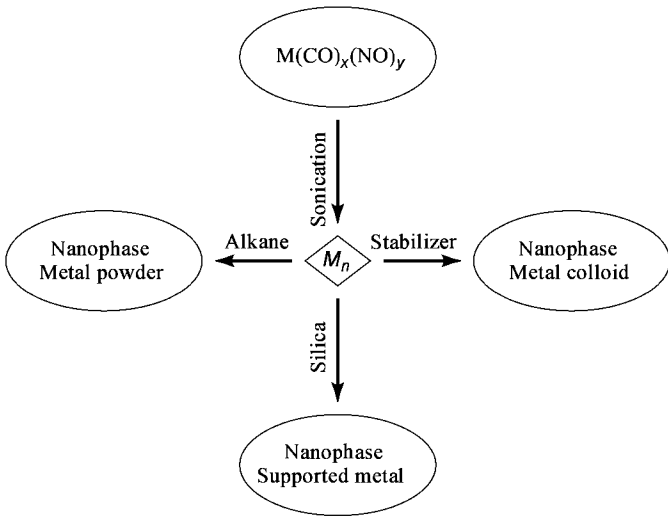


Fig. 11. Sonochemical synthesis of various forms of nanostructured materials. $n = 100-1000$.

Sonochemistry is also proving to have important applications with polymeric materials. Substantial work has been accomplished in the sonochemical initiation of polymerization and in the modification of polymers after synthesis (3,5). The use of sonolysis to create radicals which function as radical initiators has been well explored. Similarly the use of sonochemically prepared radicals and other reactive species to modify the surface properties of polymers is being developed, particularly by G. Price. Other effects of ultrasound on long chain polymers tend to be mechanical cleavage, which produces relatively uniform size distributions of shorter chain lengths.

Another important application has been the sonochemical preparation of biomaterials, most notably protein microspheres (65–68). Using high intensity ultrasound and simple protein solutions, a remarkably easy method to make both air-filled microbubbles and nonaqueous liquid-filled microcapsules has been developed. Figure 12 shows an electron micrograph of sonochemically prepared microspheres. These microspheres are stable for months, and being slightly smaller than erythrocytes, can be intravenously injected to pass unimpeded through the circulatory system. The mechanism responsible for microsphere formation is a combination of *two* acoustic phenomena: emulsification and cavitation. Ultrasonic emulsification creates the microscopic dispersion of the protein solution necessary to form the proteinaceous microspheres. Alone, however, emulsification is insufficient to produce long-lived microspheres. The long life of these microspheres comes from a sonochemical cross-linking of the protein shell. Protein cysteine residues are oxidized during microsphere formation by sonochemically produced superoxide. These protein microspheres, have a wide range of biomedical applications, including their use as echo contrast agents for sonography, magnetic-resonance-imaging contrast enhancement, drug delivery, among others.

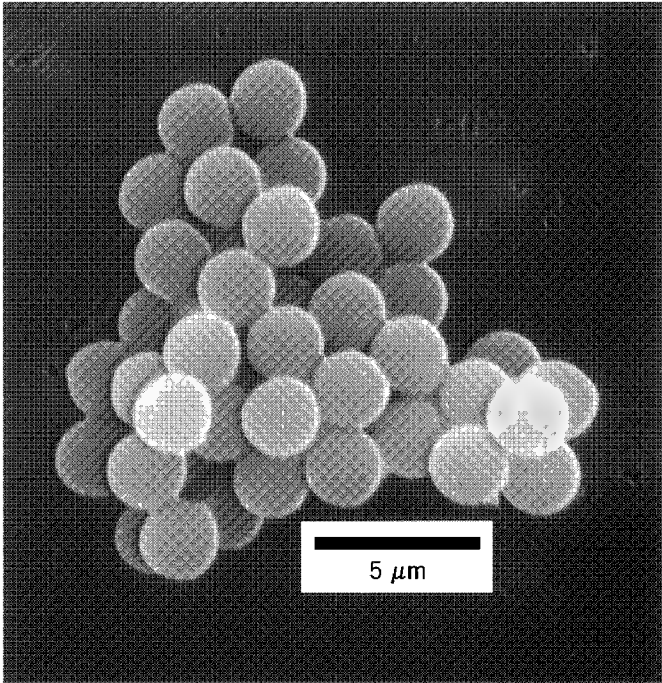


Fig. 12. Scanning electron micrograph of sonochemically synthesized hemoglobin microspheres.

Heterogeneous Sonochemistry: Reactions of Solids with Liquids. The use of ultrasound to accelerate chemical reactions in heterogeneous systems has become increasingly widespread. The physical phenomena which are responsible include the creation of emulsions at liquid–liquid interfaces, the generation of cavitational erosion and cleaning at liquid–solid interfaces, the production of shock wave damage and deformation of solid surfaces, the enhancement in surface area from fragmentation of friable solids, and the improvement of mass transport from turbulent mixing and acoustic streaming.

The use of high intensity ultrasound to enhance the reactivity of reactive metals as stoichiometric reagents has become an especially routine synthetic technique for many heterogeneous organic and organometallic reactions (11–15), particularly those involving reactive metals, such as Mg, Li or Zn. This development originated from the early work of Renaud and the more recent breakthroughs of Luche (12,13). The effects are quite general and apply to reactive inorganic salts and to main group reagents as well (69). Less work has been done with unreactive metals (eg, V, Nb, Mo, W), but results here are promising as well (11). Rate enhancements of more than tenfold are common, yields are often substantially improved, and by-products avoided. A wide range of synthetically useful heterogeneous sonochemical reactions have been listed in Table 1. The applications of sonochemistry to organic synthesis have been reviewed recently in great detail (13).

Table 1. Some Representative Examples of Heterogeneous Sonochemistry

Heterogeneous reagent	Reactant	Products
Compounds of metals		
LiAlH ₄	Ar-X	ArH
LiAlH ₄	R ₃ M-X (X = Cl, NR ₂ , OR)	R ₃ M-H (M = Si, Ge, Sn)
Al ₂ O ₃	C ₆ H ₅ CH ₂ Br + KCN	C ₆ H ₅ CH ₂ CN
KMnO ₄	RR'HCOH	RR'CO
CrO ₂ Cl ₂	RR'HCOH	RR'CO
HBR ₂	R' ₂ C=CR' ₂	HR' ₂ C-CR' ₂ (BR ₂)
MS ₂ , M = Mo, Ta, Zr	NR ₃ , py, Cp ₂ Co	intercalates
Hg	(HR ₂ BrC') ₂ CO + R'CO ₂ H	(HR ₂ C)CO(C(O ₂ CR')R ₂)
	(HR ₂ BrC') ₂ CO + R'OH	(HR ₂ C)CO(C(OR')R ₂)
Mg	R-Br	R-MgBr
	R ₂ C=CHCH ₂ Cl + Mg C ₁₄ H ₁₄	R ₂ C=CHCH ₂ MgCl
Li	R-Br (R = C ₃ H ₆ , <i>m</i> -C ₄ H ₉ , C ₆ H ₅)	R-Li
	R-Br + R'R''CO	RR'R'COH
	R-BR + (H ₃ C) ₂ NCHO	RCHO
	R ₃ M-Cl (M = C, Si, Sn)	R ₃ MMR ₃
	R ₂ SiCl ₂ (R = arenes)	cyclo-(R ₂ Si) ₃
	R ₂ MCl ₂ + Na + Se(M = Si, Sn)	(R ₂ MSe) ₃
Zn	CF ₃ I + RR'C=O	RR'C(OH)CF ₃
	C _n F _{2n+1} I + CO ₂	C _n F _{2n+1} CO ₂ H
	RR'C=O + BrCH ₂ CO ₂ R''	RR'C(OH)CH ₂ CO ₂ R''
	C ₆ H ₅ Br + RCOCH=CHR' + Ni(acac) ₂	RCOCH ₂ CHR'C ₆ H ₅ (R = H, alkyl)
	RR'C=O + R''C=CHCH ₂ Br	RR'(HO)CCR ₂ CH=CH ₂
Transition metals		
	MCl ₅ + Na + CO(M = V, Nb, Ta)	M(CO) ₆
	MCl ₆ + Na + CO(M = Cr, Mo, W)	M ₂ (CO) ₁₀ ²
	MnCl ₃ + Na + CO	Mn(CO) ₅
	FeCl ₃ + Na + CO	Fe(CO) ₄ ² + Fe ₂ (CO) ₈ ²
	Fe ₂ (CO) ₉ + alkenylepoide	Fe(CO) ₃ (π-allyllactone)
	Fe ₂ (CO) ₉ + RHC=CR'-CH=CHR'	Fe(CO) ₃ (η ⁴ -diene)
	NiCl ₂ + Na + CO	Ni ₆ (CO) ₁₂ ²
	NiCl ₂ + Na + bipy + COD CDT	Ni(bipy)(COD CDT)
	Co(acac) ₃ + C ₆ H ₆ + COD + Mg C ₁₄ H ₁₀	Co(Cp)(COD)

$\text{RuCl}_3 + \text{Li} + 1,5\text{-cyclooctadiene}$	$(\eta^{-1,3,5}\text{-COT}(\eta^4\text{-1,5-COD})\text{Ru}(0)$
$[\text{Fe}(\text{C}_5\text{R}_5(\text{CO})_2)_2 + \text{K} + (\text{C}_6\text{H}_5)_3\text{C}^+$	$\text{Fe}(\text{C}_5\text{R}_5(\text{CO})_2(\text{C}_2\text{H}_4)^+$
$[\text{M}(\text{C}_5\text{R}_5)(\text{CO})_n]_2 + \text{K} (\text{M} = \text{Fe, Ru, Mo})$	$[\text{M}(\text{C}_5\text{R}_5)(\text{CO})_n]$
$\text{Pd} + \text{CH}_2=\text{CHCH}_2\text{X}$	$[(\eta^3\text{-H}_2\text{C-CH-CH}_2)\text{Pd}(\text{X})]_2$
$\text{La} + \text{NH}_4\text{SCN in HMPA}$	$\text{La}(\text{NCS})_3\cdot 4\text{HMPA}$
$\text{Cu} + o\text{-C}_6\text{H}_4(\text{NO}_2)\text{I}$	$o\text{-(O}_2\text{N)H}_4\text{C-C H}_4(\text{NO}_2)$

The mechanism of the sonochemical rate enhancements in both stoichiometric and catalytic reactions of metals is associated with dramatic changes in morphology of both large extended surfaces and of powders. As discussed earlier, these changes originate from microjet impact on large surfaces and high velocity interparticle collisions in slurries. Surface composition studies by Auger electron spectroscopy and sputtered neutral mass spectrometry reveal that ultrasonic irradiation effectively removes surface oxide and other contaminating coatings (11). The removal of such passivating coatings can dramatically improve reaction rates. The reactivity of clean metal surfaces also appears to be responsible for the greater tendency for heterogeneous sonochemical reactions to involve single electron transfer rather than acid–base chemistry (70).

Applications of ultrasound to electrochemistry have also seen substantial recent progress. Beneficial effects of ultrasound on electroplating and on organic synthetic applications of organic electrochemistry (71) have been known for quite some time. More recent studies have focused on the underlying physical theory of enhanced mass transport near electrode surfaces (72,73). Another important application for sonoelectrochemistry has been developed by J. Reisse and co-workers for the electroreductive synthesis of submicrometer powders of transition metals (74).

Sonocatalysis. Ultrasound has potentially important applications in both homogeneous and heterogeneous catalytic systems. The inherent advantages of sonocatalysis include (1) the use of low ambient temperatures to preserve thermally sensitive substrates and to enhance selectivity; (2) the ability to generate high energy species difficult to obtain from photolysis or simple pyrolysis; and (3) the mimicry of high temperature and pressure conditions on a microscopic scale.

Homogeneous catalysis of various reactions often uses organometallic compounds. The starting organometallic compound, however, is often catalytically inactive until loss of metal-bonded ligands (such as carbon monoxide) from the metal. Having demonstrated that ultrasound can induce ligand dissociation, the initiation of homogeneous catalysis by ultrasound becomes practical. A variety of metal carbonyls upon sonication will catalyze the isomerization of 1-alkenes to the internal alkenes (57), through reversible hydrogen atom abstraction, with rate enhancements of as much as 10⁵ over thermal controls.

Heterogeneous catalysis is generally more industrially important than homogeneous systems, and the applications of ultrasound here have been reviewed (75). Heterogeneous catalysts often require rare and expensive metals. The use of ultrasound offers some hope of activating less reactive, but also less costly, metals. Such effects can occur in three distinct stages: (1) during the formation of supported catalysts, (2) activation of preformed catalysts, or (3) enhancement of catalytic behavior during a catalytic reaction. Some early investigations of the effects of ultrasound on heterogeneous catalysis can be found in the Soviet literature (76). In this early work, increases in turnover rates were usually observed upon ultrasonic irradiation, but were rarely more than tenfold. In the cases of modest rate increases, it appears likely that the cause is increased effective surface area; this is especially important in the case of catalysts supported on brittle solids (77). More impressive accelerations, however, have include hydrogenations and hydrosilations by Ni powder, Raney Ni, and Pd or Pt on carbon. For example, as shown in Figure 13, the hydrogenation of alkenes by Ni powder is enormously enhanced (>10⁵-fold) by ultrasonic irradiation (16). This dramatic increase in catalytic activity is due to the formation of uncontaminated metal surfaces from interparticle collisions caused by cavitation-induced shockwaves.

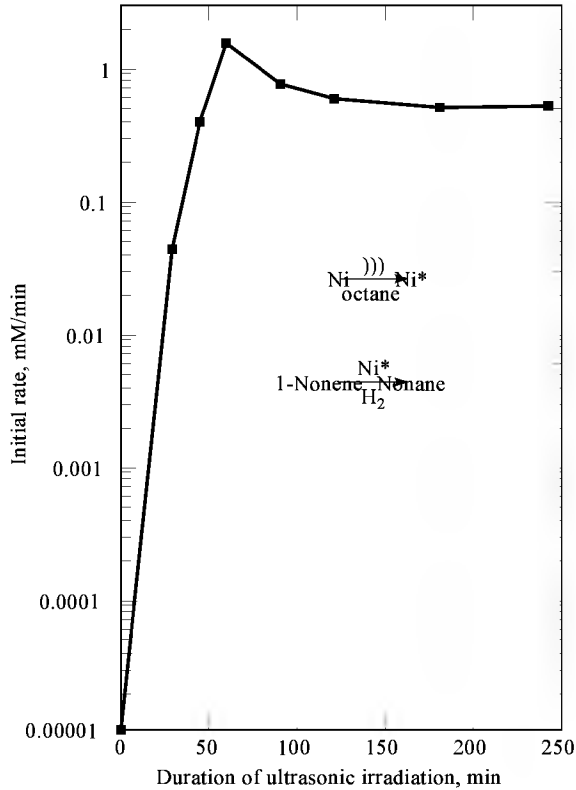


Fig. 13. The effect of ultrasonic irradiation on the catalytic hydrogenation activity of Ni powder.

Reproduced with permission (16).

Summary

The phenomenon of acoustic cavitation results in an enormous concentration of energy. If one considers the energy density in an acoustic field that produces cavitation and that in the collapsed cavitation bubble, there is an amplification factor of over eleven orders of magnitude. The enormous local temperatures and pressures so created result in phenomena such as sonochemistry and sonoluminescence and provide a unique means for fundamental studies of chemistry and physics under extreme conditions. A diverse set of applications of ultrasound to enhancing chemical reactivity has been explored, with important applications in mixed-phase synthesis, materials chemistry, and biomedical uses.

BIBLIOGRAPHY

1. K. S. Suslick, ed., *Ultrasound: Its Chemical, Physical, and Biological Effects*, VCH Publishers, New York, 1988.

2. K. S. Suslick, *Science* **247**, 1439 (1990).

3. T. J. Mason, ed., *Advances in Sonochemistry*, vols. 1–4, JAI Press, New York, 1990, 1991, 1993, 1996.

4. T. J. Mason and J. P. Lorimer, *Sonochemistry: Theory, Applications and Uses of Ultrasound in Chemistry*, Ellis Horword, Ltd., Chichester, U.K., 1988.

5. G. J. Price, ed., *Current Trends in Sonochemistry*, Royal Society of Chemistry, Cambridge, 1992.

6. K. S. Suslick, D. A. Hammerton, and R. E. Cline, Jr., *J. Am. Chem. Soc.* **108**, 5641 (1986).

7. E. B. Flint and K. S. Suslick, *Science* **253**, 1397 (1991).

8. L. A. Crum, *Proc. 1982 Ultrasonics Symp.* **1**, 1 (1982).

9. T. G. Leighton, *The Acoustic Bubble*, Academic Press, London, 1994, pp. 531–551.

10. S. J. Doktyez and K. S. Suslick, *Science* **247**, 1067 (1990).

11. K. S. Suslick and S. J. Doktycz, *Adv. Sonochem.* **1**, 197–230 (1990).

12. C. Einhorn, J. Einhorn, and J.-L. Luche, *Synthesis*, 787 (1989).

13. J. L. Luche, *Comptes Rendus Serie IIB* **323**, 203, 337 (1996).

14. J. M. Pestman, J. B. F. N. Engberts, and F. de Jong, *Recl. Trav. Chim. Pays-Bas* **113**, 533 (1994).

15. K. S. Suslick, "Sonochemistry of Transition Metal Compounds," in R. B. King, ed., *Encyclopedia of Inorganic Chemistry*, John Wiley & Sons, Inc., New York, vol. 7, pp. 3890–3905.

16. K. S. Suslick and D. J. Casadonte, *J. Am. Chem. Soc.* **109**, 3459 (1987).

17. L. A. Crum, *Physics Today* **47**, 22 (1994).

18. S. J. Putterman, *Scientific American*, 46 (Feb. 1995).

19. Lord Rayleigh, *Philos. Mag.* **34**, 94 (1917).

20. W. T. Richards and A. L. Loomis, *J. Am. Chem. Soc.* **49**, 3086 (1927).

21. M. A. Margulis, *Ultrasonics* **30**, 152 (1992).

22. T. Lepoint and F. Mullie, *Ultrasonics Sonochem.* **1**, S13 (1994).

23. H. G. Flynn, "Physics of Acoustic Cavitation in Liquids," in W. P. Mason, ed., *Physical Acoustics*, vol. 1B, Academic Press, New York, 1964, p. 157.

24. L. A. Crum, *J. Acoust. Soc. Am.* **95**, 559 (1994).

25. B. P. Barber and S. J. Putterman, *Phys. Rev. Lett.* **69**, 3839 (1992).

26. R. Lofstedt, B. P. Barber, and S. J. Putterman, *Phys. Fluids A* **5**, 2911 (1993).

27. A. Henglein, *Ultrasonics* **25**, 6 (1985).

28. A. Henglein, *Adv. Sonochem.* **3**, 17 (1993).

29. H. Frenzel and H. Schultes, *Z. Phys. Chem.* **27b**, 421 (1934).

30. T. J. Matula, R. A. Roy, P. D. Mourad, W. B. McNamara III, and K. S. Suslick, *Phys. Rev. Lett.* **75**, 2602 (1995).

31. D. F. Gaitan and L. A. Crum, in D. T. Blackstock, ed., *Frontiers of Nonlinear Acoustics*, 12th ISNA, Elsevier Applied Science, New York, 1990, pp. 459–463.

32. D. F. Gaitan, L. A. Crum, R. A. Roy, and C. C. Church, *J. Acoust. Soc. Am.* **19**, 3166 (1992).

33. B. P. Barber, R. A. Hiller, R. Lofstedt, S. J. Putterman, and K. R. Weninger, *Phys. Rev. Lett.* **72**, 1380 (1994).

34. L. A. Crum and R. A. Roy, *Science* **266**, 233 (1994).

35. R. E. Verrall and C. Sehgal in Ref. 1, pp. 227–287.

36. E. B. Flint and K. S. Suslick, *J. Phys. Chem.* **95**, 1484 (1991).

37. E. B. Flint and K. S. Suslick, *J. Amer. Chem. Soc.* **111**, 6987 (1989).

38. R. Hiller, K. Weninger, S. J. Putterman, and B. P. Barber, *Science* **266**, 248 (1994).

39. B. P. Barber, R. Hiller, K. Arisaka, H. Fetterman, and S. J. Putterman, *J. Acoust. Soc. Am.* **91**, 3061 (1992).

40. B. P. Barber, C. C. Wu, R. Lofstedt, P. H. Roberts, and S. J. Putterman, *Phys. Rev. Lett.* **72**, 1380 (1994).

41. W. C. Moss, D. B. Clarke, and D. A. Young, *Science* **276**, 1398 (1997).

42. D. Lohse and co-workers, *Phys. Rev. Lett.* **78**, 1359 (1997).

43. J. Bedan and T. J. Mason, *Ultrasonics* **30**, 203 (1992).

44. T. J. Mason and E. D. Cordemans, *Chem. Eng. Res. Des.* **74**, 511 (1996).

45. R. L. Hunicke, *Ultrasonics* **28**, 291 (1990).

46. A. Shoh in Ref. 1, pp. 97–122.

47. J. Reisse, T. Caulier, C. Deckerkheer, O. Fabre, J. Vandercammen, J. L. Delplancke, and R. Winand, *Ultrason. Sonochem.* **3**, S147 (1996).

48. C. Petrier and S. Laguian, *Chemosphere* **32**, 1709 (1996).

49. K. S. Suslick, J. W. Gawienowski, P. F. Schubert, and H. H. Wang, *Ultrasonics* **22**, 33 (1984).

50. K. S. Suslick, J. W. Gawienowski, P. F. Schubert, and H. H. Wang, *J. Phys. Chem.* **87**, 2229 (1983).

51. P. Riez, *Adv. Sonochem.* **2**, 23 (1991).

52. V. Misik and P. Riesz, *Ultrason. Sonochem.* **3**, S173 (1996).

53. P. Riesz, D. Berdahl, and C. L. Christman, *Environ. Health Perspect.* **64**, 233 (1985).

54. M. A. Margulis and N. A. Maximenko, *Adv. Sonochem.* **2**, 253 (1991).

55. M. R. Hoffmann, I. Hua, and R. Hochemer, *Ultrasonics Sonochemistry* **3**, S163 (1996).

56. I. Hua, R. H. Hochemer, and M. R. Hoffmann, *Env. Sci. Tech.* **29**, 2790 (1995).

57. K. S. Suslick, *Adv. Organomet. Chem.* **25**, 73 (1986).

58. K. S. Suslick, *MRS Bull.* **20**, 29 (1995).

59. O. V. Abramov, *Ultrasound in Liquid and Solid Metals*, CRC Press, Boca Raton, Fla., 1994.

60. K. S. Suslick, S. B. Choe, A. A. Cichowlas, and M. W. Grinstaff, *Nature* **353**, 414 (1991).

61. M. W. Grinstaff, M. B. Salamon, and K. S. Suslick, *Phys. Rev. B* **48**, 269 (1993).

62. K. S. Suslick, T. Hyeon, and M. Fang, *Chem. Mater.* **8**, 2172 (1996).

63. T. Hyeon, M. Fang, and K. S. Suslick, *J. Am. Chem. Soc.* **118**, 5492 (1996).

64. K. S. Suslick, M. Fang, and T. Hyeon, *J. Am. Chem. Soc.* **118**, 11960 (1996).

65. K. S. Suslick and M. W. Grinstaff, *J. Am. Chem. Soc.* **112**, 7807 (1990).

66. K. S. Suslick and M. W. Grinstaff, *Proc. Natl. Acad. Sci. USA* **88**, 7708 (1991).

67. K. J. Liu, M. W. Grinstaff, J. Jiang, K. S. Suslick, H. M. Swartz, and W. Wang, *Biophys. J.* **67**, 896 (1994).

68. A. G. Webb, M. Wong, K. J. Kolbeck, R. L. Magin, L. J. Wilmes, and K. S. Suslick, *J. Mag. Res. Imaging* **6**, 675 (1996).

69. T. Ando and T. Kimura, *Adv. Sonochem.* **2**, 211 (1991).

70. J.-L. Luche, *Ultrasonics Sonochem.* **1**, S111 (1994).

71. A. Durant, H. Francois, J. Reisse, and A. Kirschdemesmaeker, *Electrochim. Acta* **41**, 277 (1996).

72. R. G. Compton, J. C. Eklund, and F. Marken, *Electroanalysis*. **9**, 509 (1997).

73. N. A. Madigan and L. A. Coury, *Anal. Chem.* **69**, 5 (1997).

74. A. Durant, J. L. Delplancke, R. Winand, and J. Reisse, *Tetrahedron Lett.* **36**, 4257 (1995).
75. K. S. Suslick, "Sonocatalysis," in G. Ertl, H. Knozinger, and J. Weitkamp, eds., *Handbook of Heterogeneous Catalysis*, vol. 3, Wiley-VCH, Weinheim, 1997, Chapt. 8.6, pp. 1350–1357.
76. A. N. Mal'tsev, *Zh. Fiz. Khim.* **50**, 1641 (1976).
77. B. H. Han and P. Boudjouk, *Organometallics* **2**, 769 (1983).

Kenneth S. Suslick
University of Illinois at Urbana-Champaign

SURFACE AND INTERFACE ANALYSIS

Surfaces are formed in the transition from one state of matter to another, whether the two phases are chemically distinct or not. Thus, surfaces exist at interphases or interfaces between two phases of either the same or different materials. For example, the surface of an ice cube in a glass of water represents an interface between two phases that are identical in chemical composition. The surface of a straw in the same glass of water represents an example of an interface between chemically distinct materials.

Many important chemical processes occur at solid–solid, solid–liquid, solid–gas, liquid–liquid, and liquid–gas interfaces. For example, many catalytic processes such as the conversion of petroleum to high octane fuels are routinely performed at the surface of metal oxides. These surfaces can act either to provide sites enhancing proximity of reactants or mechanistically be involved in the chemical transformation itself. Environmental concerns often center on the extent to which organic pollutants are transported through solid soil, clay or mineral surfaces; organic pollutants that strongly interact with these surfaces will be transported less efficiently than those which do not experience such interactions. Surfactant molecules are essential ingredients in soaps, foams and emulsions; the unique chemistry of these systems arises from the formation of interfaces between two liquids or between a liquid and a gas. Thus, the ability to characterize surfaces and interfaces is essential to better understanding a variety of technologically and societally important chemical processes.

Overview of Surface and Interface Analysis Methods

A plethora of characteristics of surfaces and interfaces can be analyzed using a staggering variety of techniques. The tools of surface and interface analysis have been undergoing rapid development during the past several decades. As significant advances in instrumentation are realized, new and improved methods for surface and interface analysis will emerge. Given the scope of modern surface and interfacial analysis, all of the methods used for study cannot be adequately treated here. Therefore, the focus of this article will be on those methods that are deemed by the author to be of the greatest importance. The interested reader is referred to numerous other compendiums of information on this broad topic (1–5). Particularly noteworthy are the series of Fundamental Reviews on specific techniques that appear biennially in the journal *Analytical Chemistry*. These Reviews report developments in specific fields since the previous report, and usually provide an up-to-date perspective of significant advances made in the field. Of particular relevance to this article are the Fundamental Reviews on surface analysis, scanning probe microscopy, and ir spectroscopy which have appeared recently (5).

The chemical, structural, and electronic characteristics of surfaces and interfaces are usually different from those of the bulk phase(s). Thus, methods to be used for the analysis of surfaces must be selective in response to the surface or interfacial region relative to the bulk. Surfaces and interfaces are most commonly explored using techniques based on the interaction of photons, electrons, or ions with the surface or using a force such as electric field or van der Waals attraction. These excitations generate a response involving the production of photons, electrons, ions or the alteration of a force that is then sensed in the analysis.

The choice of which of these exciting species or forces, and hence technique, to use depends on the nature of the information sought about the surface. This article is organized around a discussion of techniques which can be used to answer the following important fundamental questions that are typically asked about surfaces and interfaces:

- What does the surface or interface look like?
- What is the elemental composition of the surface or interface?
- What molecules are at the surface or in the interface?

Table 1 provides an overview of many of the techniques available for the characterization of surfaces and interfaces. These techniques are categorized on the basis of the nature of the exciting and detected species (or force). As can be seen by Table 1, a tremendous number of approaches are available for the study of surfaces. In fact, multiple methods capable of answering all of the three questions posed above have been developed over the past thirty years.

Table 1. Overview of Common Surface Analysis Techniques

Excitation	Detection			
	Photons	Electrons	Ions	Force
photons	ellipsometry	xps ^k (esca ^l)	psd ⁿ	
	ftir ^a , irras ^b , drift ^c	ups ^m		
	Raman, sers ^d			
	shg ^e , sfg ^f			
	xrd ^g			
	exafs ^h , nexafs ⁱ , sexafs ^j			
electrons	ip ^o	sem ^p , fe-sem ^q , tem ^r , stem ^s	esdiad ^{aa}	
		aes ^t , aps ^u		
		leed ^v , rheed ^w		
		eels ^x , hreels ^y , ceels ^z		
ions		ins ^{bb}	sims ^{cc}	
			iss ^{dd} , rbs ^{ee}	
			leis ^{ff}	
force		fem ^{gg}	fim ^{hh}	afm ⁱⁱ
		stm ^{jj}		

^a ftir = Fourier transform infrared spectroscopy.
^b irras = infrared reflection absorption spectroscopy.
^c drift = diffuse reflectance infrared Fourier transform spectroscopy.
^d sers = surface enhanced Raman scattering.
^e shg = second harmonic generation.

- ^f sfg = sum-frequency generation.
- ^g xrd = x-ray diffraction.
- ^h exafs = extended x-ray absorption fine structure.
- ⁱ nexafs = near-edge x-ray absorption fine structure.
- ^l sexafs = surface extended x-ray absorption fine structure.
- ^k xps = x-ray photoelectron spectroscopy.
- ^l esca = electron spectroscopy for chemical analysis.
- ^m ups = ultraviolet photoelectron spectroscopy.
- ⁿ psd = photon stimulated ion angular distribution.
- ^o ip = inverse photoemission.
- ^p sem = scanning (or secondary) electron microscopy.
- ^q fe-sem = field emission scanning electron microscopy.
- ^r tem = transmission electron microscopy.
- ^s stem = scanning transmission electron microscopy.
- ^t aes = Auger electron spectroscopy.
- ^u aps = appearance potential spectroscopy.
- ^v leed = low energy electron diffraction.
- ^w rheed = reflected high energy electron diffraction.
- ^x eels = electron energy loss spectroscopy.
- ^y hreels = high resolution electron energy loss spectroscopy.
- ^z ceels = core level electron energy loss spectroscopy.
- ^{aa} esdiad = electron-stimulated ion angular distribution.
- ^{bb} ins = ion neutralization spectroscopy.
- ^{cc} sims = secondary ion mass spectroscopy.
- ^{dd} iss = ion scattering spectroscopy.
- ^{ee} rbs = Rutherford backscattering spectroscopy.
- ^{ff} leis = low energy ion scattering.
- ^{gg} fem = field emission microscopy.
- ^{hh} fim = field ion microscopy.
- ⁱⁱ afm = atomic force microscopy.
- ⁱⁱ stm = scanning tunneling microscopy.

Several methods utilizing photons as the exciting and detected species are commonly employed. Ellipsometry provides information about thin-film thickness, typically using visible wavelength light. Information about the vibrational modes of molecules at surfaces or within interfaces can be obtained using one of several vibrational spectroscopies including transmission ftr spectroscopy, diffuse reflectance infrared Fourier transform spectroscopy (drtf), infrared reflection-absorption spectroscopy (irras), Raman spectroscopy, surface enhanced Raman scattering (sers), or sum-frequency generation (sfg). These approaches are based on visible or infrared wavelength photons that are either absorbed, scattered, or generated. X-ray diffraction (xrd) is a surface structural technique that relies on the diffraction of x-ray wavelength photons. Extended x-ray absorption fine structure (exafs) is a technique in which the x-ray absorption cross-section of a material is modulated by the local surface structure and coordination environment of an atom. This modulation is detected by either x-ray fluorescence or attenuation of the incident x-ray beam. In versions of this technique which are more surface sensitive, near-edge x-ray absorption fine structure (nexafs) or surface extended x-ray absorption fine structure (sexafs), x-ray absorption is modulated by surface structure as sensed by monitoring photoemitted or Auger electrons.

Other techniques in which incident photons excite the surface to produce detected electrons are also listed in Table 1. X-ray photoelectron spectroscopy (xps), which is also known as electron spectroscopy for chemical analysis (esca), is based on the use of x-rays which stimulate atomic core level electron ejection for elemental composition information. Ultraviolet photoelectron spectroscopy (ups) is similar but uses ultraviolet photons instead of x-rays to probe atomic valence level electrons. Photons are used to stimulate desorption of ions in photon stimulated ion angular distribution (psd). Inverse photoemission (ip) occurs when electrons incident on a surface result in photon emission which is then detected.

Electrons are used for excitation and detection in the various electron microscopy methods for surface visualization and imaging [scanning electron microscopy (sem), field emission scanning electron microscopy (fe-sem), transmission electron microscopy (tem), and scanning transmission electron microscopy (stem)], the electron diffraction methods for the determination of surface atomic structure [low energy electron diffraction (leed), reflected high energy electron diffraction (rheed)], and the electron energy loss spectroscopies for evaluation of surface electronic and chemical characteristics [electron energy loss spectroscopy (eels), high resolution electron energy loss spectroscopy (hreels), and core level electron energy loss spectroscopy (ceels)]. Ions produced by electron-stimulated desorption are measured in electron-stimulated desorption ion angular distribution (esdiad). If neutralization of an incident ion beam is measured by surface electron ejection, the technique of ion neutralization spectroscopy (ins) results.

Ions can also be used as both the excitation and detection species in several surface analysis techniques. In secondary ion mass spectroscopy (sims), an incident beam of ions can sputter molecular ions from a surface providing surface elemental and molecular information. Ions can be scattered from surfaces or within interfaces in ion scattering spectroscopy (iss), Rutherford backscattering spectroscopy (rbs), and low energy ion scattering (leis) to provide elemental composition and structural information. Finally, an electric field can be used to stimulate ionization of an imaging gas in field ion microscopy (fim) or electron tunneling in field emission microscopy (fem) and scanning tunneling microscopy (stm) for surface imaging. A surface imaging technique which utilizes van der Waals forces as the "excitation" while monitoring what are essentially drag or frictional forces is atomic force microscopy (afm).

In choosing appropriate techniques with which to interrogate surfaces in search of answers to the above three questions, one must remain cognizant of the lateral resolution and depth at which the chemistry of interest occurs as well as the lateral resolution and depth sensitivity of each technique. Photon beams cannot be focused as tightly as electron or ion beams, for example, implying that techniques using photon beams will have poorer lateral resolution than techniques using electron beams. This characteristic may or may not be a limitation, however, depending on the scale on which the surface chemistry of interest is occurring.

Depth sensitivity is an equally important consideration in the analysis of surfaces. Techniques based on the detection of electrons or ions derive their surface sensitivity from the fact that these species cannot travel long distances in solids without undergoing interactions which cause energy loss. If electrons are used as the basis of an analysis, the depth resolution will be relatively shallow and depend on both the energy of the incident and detected electrons and on characteristics of the material. In contrast, techniques based on high energy photons such as x-rays will sample a much greater depth due

to the fact that x-rays can travel long distances in solids. Therefore, the depth sensitivity of a particular method will depend on the nature and energy of both the probe and detected species, and may or may not be sensitive to the entire depth to which the chemistry of interest occurs.

Although not a universal requirement of surface analysis, techniques that utilize electron or ion beams must be carried out in ultrahigh vacuum (UHV) environments wherein the pressures are kept lower than $ca\ 10^{-8}$ Pa so that these species can travel undisturbed until they impinge on the surface of interest. Vacuum technology is significantly advanced such that these pressures can be routinely achieved and maintained as a matter of course with adequate pumping in usually rigorously clean, leak-free, stainless steel environments. A variety of pumps exist which can pump down to UHV pressures with turbomolecular, ion, and diffusion pumps being the most common. Each of these pumps possesses its own advantages and disadvantages. The interested reader is referred to Reference 6 for a more comprehensive treatment of modern vacuum technology.

Techniques

Imaging of Surfaces—Analysis of Surface Morphology. Several important techniques can help answer the question: what does the surface look like? This question is often the first one to be posed in the characterization of a new surface or interface. Physical imaging of the surface is necessary to distinguish the relevant features important for understanding the whole surface and is essential for accurate interpretation of data from other surface analysis techniques which might later be applied to a more limited region of the surface or interface.

Electron Microscopy. The techniques of primary importance in surface imaging are the electron microscopies (7–13). These techniques take advantage of the fact that accelerated beams of electrons typically have higher energies than commonly available photon beams, and therefore, can be focused more tightly. For example, consider two electron beams of 10 keV and 40 keV energy, respectively. Using the fact that $E = hc/\lambda$, the wavelength λ of the electrons in each beam are calculated to be 0.124 and 0.03 nm (1.24 and 0.31 Å), respectively, much shorter than even x-ray beams. For surface imaging applications, the consequence of this wavelength is realized in terms of the lateral resolution that can be attained. Assuming that these beams are perfectly focused, the theoretically best resolution that can be achieved is on the order of $\lambda/2$ or $ca\ 0.06$ nm for the 10 keV beam and 0.015 nm for the 40 keV beam. Although this resolution is never attainable in reality due to limitations of electron lenses, these calculations serve to demonstrate the importance of using high energy beams for physical imaging on the atomic scale in electron microscopy.

Electron microscopy is done in one of two ways: scanning electron microscopy (sem) or transmission electron microscopy (tem). Scanning transmission electron microscopy (stem) is simply tem carried out in a scanning mode. Sem is the most common and well-known electron microscopy method for the physical imaging of surfaces. This technique is based on the interaction with a surface of a primary beam of electrons with energy typically in the range of 0.5–40 keV. This primary electron beam is first demagnified by a condenser lens and then focused onto the sample surface using a series of objective lenses where it is rastered across the surface using a series of scanning coils. Sem must be done in vacuum so that the electrons can travel unimpeded for adequate distances. The typical surface magnification realized is on the order of 10^3 – $10^5\times$ depending on the energy of the primary electron beam.

In most sem analyses, secondary electrons created near the position of the impinging primary beam are detected. However, surface images can also be obtained through collection of back-scattered primary electrons. The secondary electrons, which result from ionization of the sample atoms in the near-surface region by the incident or back-scattered primary electrons, have much smaller energies of $ca\ 0$ – 20 eV. Various mechanisms result in the ejection of these secondary electrons from the surface. However, due to their very low energies, secondary electrons have escape depths of only a few nm. Therefore, sem is a highly surface sensitive tool even though the primary beam electrons may penetrate into the sample to depths of 0.5 – $5\ \mu\text{m}$.

Secondary electrons can usually be collected with very high efficiencies. The efficiency of secondary electron creation, therefore, becomes important in sensitivity. The yield of secondary electrons per primary electron is dictated primarily by two factors, the primary beam energy and the angle of collection. The secondary electron yield actually shows a slight negative dependence on primary beam energy since the higher the primary beam energy, the deeper its penetration and the less efficient is the secondary electron escape from the surface. Thus, the secondary electron yield varies as $ca\ E^{-0.8}$ for primary beam energy E . In addition, the secondary electron yield depends on angle of collection, generally increasing as the angle from the surface normal increases.

Given the option to obtain images by collection of either secondary electrons or backscattered primary electrons, one might reasonably ask which is the better strategy for surface imaging. Both approaches have strengths and weaknesses depending on the information desired. Generally, images from secondary electrons have higher resolution, since they depend only on the primary beam diameter. Secondary electron images also have higher contrast than back-scattered electron images. However, backscattered electron images have their own set of advantages. Chief among these is that backscattered electron images are generally more sensitive to topography than secondary electron images, since backscattering increases significantly at sharp edges on a surface. Backscattering efficiency is a function of the atomic number (Z) of the atom from which the backscattering occurs. Thus, back-scattered images have a pronounced dependence on elemental composition of the sample surface. Secondary electron images actually have a slight dependence on elemental composition as well, since some of the secondary electrons are generated by back-scattered primary electrons. When acquiring back-scattered electron images, either of these attributes can be maximally exploited by the manner in which the collected back-scattered electrons are processed. Back-scattered electrons are detected with a two-half annular detector. Images sensitive to elemental composition differences across the surface are obtained when the electron intensities in each half of the detector are added. In contrast, images sensitive to surface topography are obtained by subtracting the electron intensities in each half of the detector.

A block diagram of a scanning electron microscope is shown in Figure 1. A vacuum on the order of $ca\ 10^{-5}$ Pa is required in order to ensure adequate electron pathlengths. This level of vacuum is readily attainable with diffusion pumps trapped with liquid nitrogen.

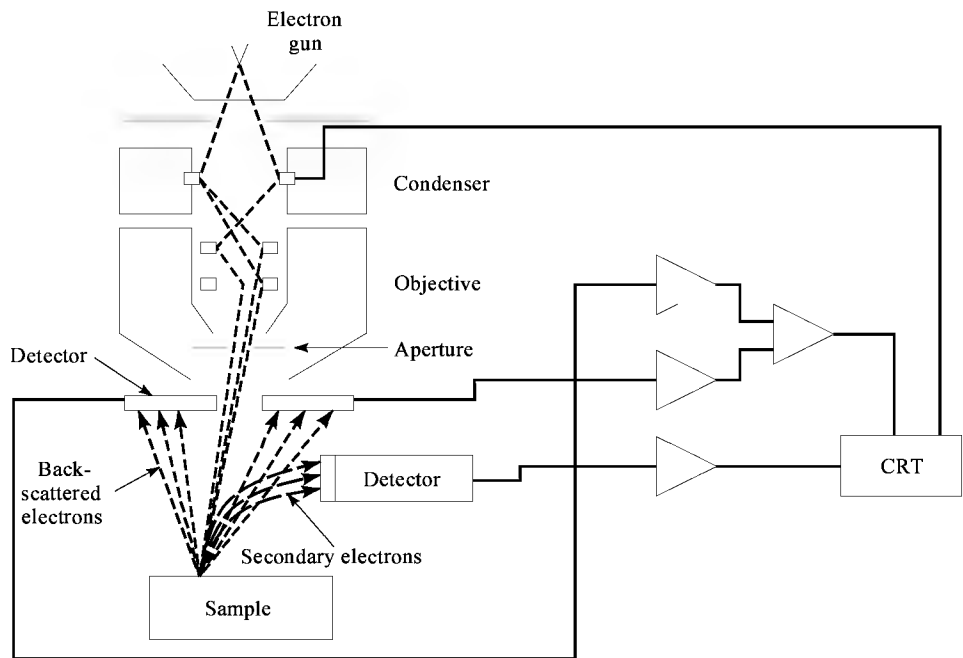


Fig. 1. Block diagram of scanning electron microscope.

The electron sources used in most sems are thermionic sources in which electrons are emitted from very hot filaments made of either tungsten (W) or lanthanum boride (LaB₆). W sources are typically heated to ca 2500–3000 K in order to achieve an adequate electron brightness. LaB₆ sources require lower temperatures to achieve the same brightness, although they need a better vacuum than W sources. Once created, these primary electrons are accelerated to some desired energy with an energy spread (which ultimately determines lateral resolution) on the order of ca 1.5 eV.

One important sem source that is not based on thermionic emission is the field emission (fe) source. Fe-sem systems typically give images of much higher resolution than conventional sems due to the much narrower energy distribution (on the order of 0.25 eV) of the primary electron beam. A fe source is a pointed W tip from which electrons tunnel under the influence of a large electric field. This different mechanism of electron generation also results in a brightness comparable to a conventional thermionic source with much less current.

Once the primary electron beam is created, it must be demagnified with condenser lenses and then focused onto the sample with objective lenses. These electron lenses are electromagnetic in nature and use electric and magnetic fields to steer the electrons. Such lenses are subject to severe spherical and chromatic aberrations. Therefore, a point primary beam source is blurred into a primary beam disk to an extent dependent on the energy and energy spread of the primary electrons. In addition, these lenses are also subject to astigmatism. All three of these effects ultimately limit the primary beam spot size and hence, the lateral resolution achievable with sem.

The detectors used in sem are of two types depending on whether secondary or back-scattered electrons are being detected. As mentioned above, back-scattered electrons are detected with an annular detector placed above the sample such that electrons back-scattered along the surface normal can be detected. Secondary electrons are detected with a detector that is positively charged to attract these electrons. This detector is placed at large angles with respect to the surface normal to most efficiently detect secondary electrons whose paths are predominantly away from the surface normal as discussed above. The absolute efficiencies of both detectors are relatively low; therefore, the signals must be amplified before being sent to an image intensifier (CRT) for viewing. For sem, the electron scan rate in the CRT is set to match the raster rate of the primary electron beam across the surface so that an image in real time is observed.

A scanning electron microscope can also be equipped with additional instrumentation for electron-excited x-ray analysis (9). In many systems, this is performed in the mode known as energy dispersive x-ray analysis (edx). Other common acronyms for this method are eds for energy dispersive spectroscopy or edax for energy dispersive analysis of x-rays.

Edx is based on the emission of x-rays with energies characteristic of the atom from which they originate in lieu of secondary electron emission. Thus, this technique can be used to provide elemental information about the sample. In the sem, this process is stimulated by the incident primary beam of electrons. As will be discussed below, this process is also the basis of essentially the same technique but performed in an electron spectrometer. When carried out this way, the technique is known as electron microprobe analysis (ema).

Sem-edx typically occurs in a volume of sample larger than that from which backscattered electrons are observed. Thus, sem-edx samples the surface to a greater depth than does sem imaging. Signals typically result from the upper several μm of the near-surface region. The intensity of the x-rays emanating from a sample at a given energy is proportional to the number density of atoms generating these x-rays. Thus, sem-edx can be used semiquantitatively. The fundamental principles and instrumentation for this approach are discussed in greater detail below in the section describing techniques for elemental composition.

A second important mode of electron microscopy is transmission electron microscopy (tem) (11). An image of a sample in tem is obtained using the transmission of electrons by a sample in a method analogous to optical microscopy using photons. Thus, this method provides a magnified image of a transparent sample with an electron beam using objective and projector lenses as shown in Figure 2. Obviously, the requirement for an electron beam-transparent sample is the most important limitation of this approach. For most materials, this means that the sample thickness must be on the order of ca 20–200 nm. Given the extremely small thickness of tem samples, the surface area-to-bulk volume is very large; therefore, tem provides information which is essentially surface in nature.

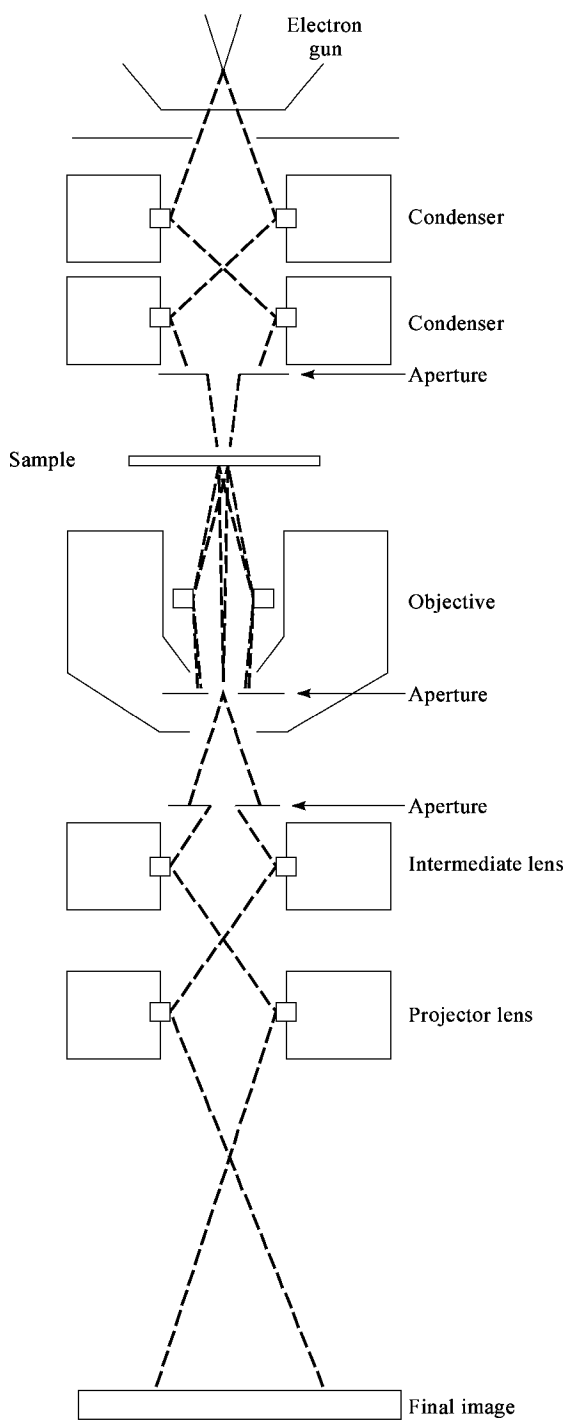


Fig. 2. Block diagram of a transmission electron microscope.

Instrumentation for tem is somewhat similar to that for sem; however, because of the need to keep the sample surface as clean as possible throughout the analysis to avoid imaging surface contamination as opposed to the sample surface itself, ultrahigh vacuum conditions (ca 10^{-7} – 10^{-8} Pa) are needed in the sample area of the microscope. Electron sources in tem are similar to those used in sem, although primary electron beam energies needed for effective tem are higher, typically on the order of ca 100 keV.

Several lenses are used in a transmission electron microscope. The condenser lenses provide uniform illumination of the sample over the area of interest. The objective lens provides the primary image and therefore, determines the lateral resolution of the image. The objective lens aperture is important in controlling the contrast of the image. The final magnification of the image is performed by one or more projector lenses. The final image is typically recorded on a fluorescent or phosphorescent screen where it can be captured by a video camera for viewing. As noted above, all of these lenses are subject to serious aberrations which ultimately limit the resolution of the microscope to greater than the diffraction limit (the theoretical resolution limit for this approach.) Moreover, these lens aberrations restrict the angular range of the electron beam resulting in the need for very tall instruments. Despite these shortcomings, tem is a very powerful surface imaging tool with atomic resolution in some cases, providing sample magnifications between 100–500,000 X.

Scanning Probe Microscopies. A relatively new group of techniques for surface imaging is the scanning probe microscopies (14,15). The most common of these tools are scanning tunneling microscopy (stm) and atomic force microscopy (afm). These techniques were developed in the 1980s and have undergone explosive growth and acceptance since their initial discovery. Stm, the first scanning probe microscopy method to be developed, was discovered by Binnig and Rohrer (16) in 1981 at IBM Zurich. These researchers received the 1986 Nobel Prize in Physics for their work. The significance of this discovery is a readily accessible technique capable of generating real-space images of a surface with atomic-scale resolution.

Stm is based on the tunneling of electrons between a very sharp, electrically conductive tip (hopefully terminating in a single atom) and a conductive sample when the tip is brought very close (<1 nm) to the sample. A schematic for this technique is shown in Figure 3. The magnitude of the tunneling current will vary in a well-defined exponential manner with tip-to-sample distance. Thus, the magnitude of the tunneling current can be used to generate an image of the surface. Stm requires that both the tip and the sample be electrically conductive. Therefore, the samples that can be studied with stm are restricted to either conductors or semiconductors. The exponential dependence of tunneling current on distance between the tip and the sample gives outstanding (ie, sub-nm) vertical resolution. Lateral resolution is not quite as good and is limited by the size of the tip.

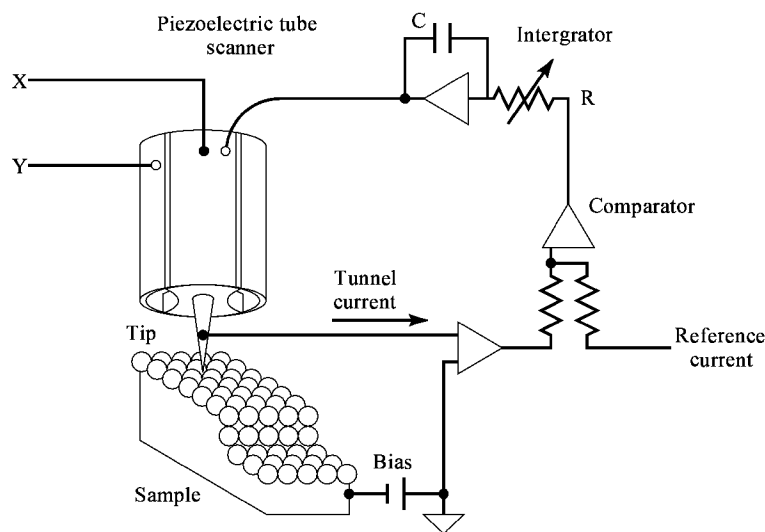


Fig. 3. Block diagram of a scanning tunneling microscope (17).

Stm experiments can be performed in one of two common operating modes as shown in Figure 4. In the constant-current mode of operation, the instrument uses electronic feedback to keep the tunneling current constant by moving the tip closer or further away from the surface as the tip is scanned across the sample. Therefore, in the simplest analysis, the height of the tip is a direct measure of the sample surface topography. This method of operation is somewhat more time-consuming than the constant height mode described below, because of the time that it takes to move the tip up and down. However, this approach is excellent for imaging irregular surfaces with high precision.

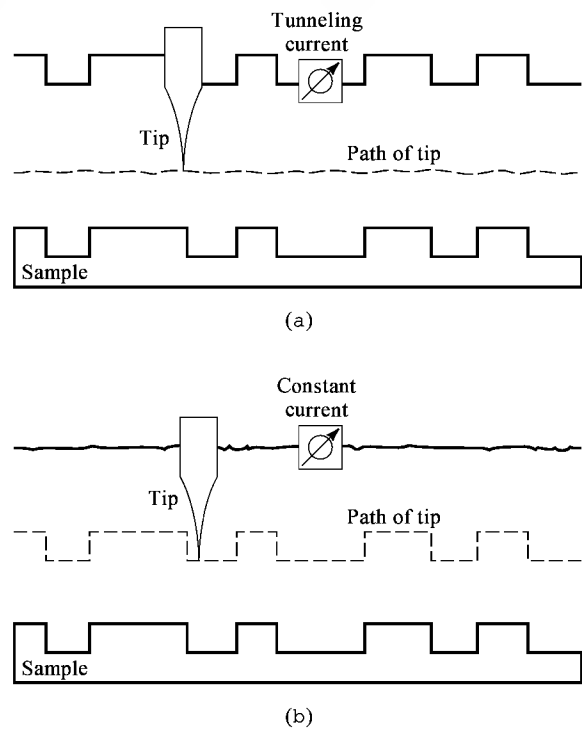


Fig. 4. Operational modes for stm. (a) Constant height mode. (b) Constant current mode (18).

The constant height mode of operation results in a faster measurement. In this analysis, the tip height is maintained at a constant level above the surface and differences in tunneling current are measured as the tip is scanned across the surface. This approach is not as sensitive to surface irregularities as the constant current mode, but it does work well for relatively smooth surfaces.

The simplest interpretation of stm images is in terms of surface topography. However, care must be exercised in this interpretation, since in reality, tunneling probability is really measured. The many subtleties of stm data interpretation are beyond the scope of this article. The interested reader is referred to references 14 and 15 for a more detailed discussion of these issues.

A second scanning probe microscopy that is perhaps more widely used for surface imaging today is afm. This technique is related to stm, but provides a direct topographical image of a surface without the complications arising from the electronic surface structure that cloud image interpretation in stm. In this device, a sharp tip several microns long terminating in a tip <10 nm in diameter is attached to a cantilever several hundred microns in length. In the original configuration of this method shown in Figure 5, the tip is dragged across the surface (or the surface dragged under the tip). Van der Waals forces between the tip and the surface cause the cantilever to bend, and the extent of this cantilever deflection is measured. The real utility of this approach when compared to stm is that it can be applied to *any* surface, whether it be insulating or conductive. Thus, this technique has found successful application to conductors, semiconductors and insulators alike.

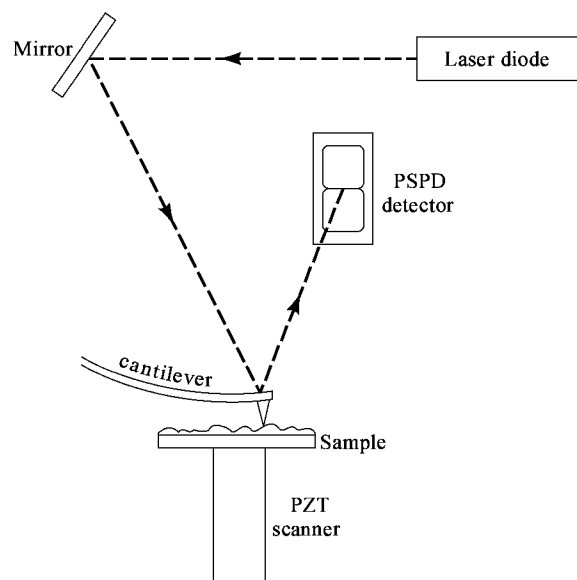


Fig. 5. Block diagram of contact atomic force microscope system in which cantilever deflection monitored optically with position-sensitive photodiode (PSPD). PZT = piezo-electric scanner (18).

Consideration of the shape of the van der Waals force versus distance curve (Fig. 6) leads to several common operational modes of afm. In contact-mode afm (see Fig. 5), the tip makes physical "contact" with the surface. This contact is close enough in distance to result in repulsive interatomic interactions between the tip and the sample. The cantilever bends as the tip is scanned across the surface due to these repulsive forces at different surface topologies. The bending of the cantilever is commonly sensed using an optical technique involving deflection of a laser beam reflected off of the cantilever as sensed by a position-sensitive photodiode. This arrangement is sensitive to <0.1 nm vertical displacement of the cantilever position. As the tip is dragged across the surface, it experiences, in addition to repulsive forces, frictional or drag forces which can be measured as well. These are the basis of frictional force microscopy (ffm).

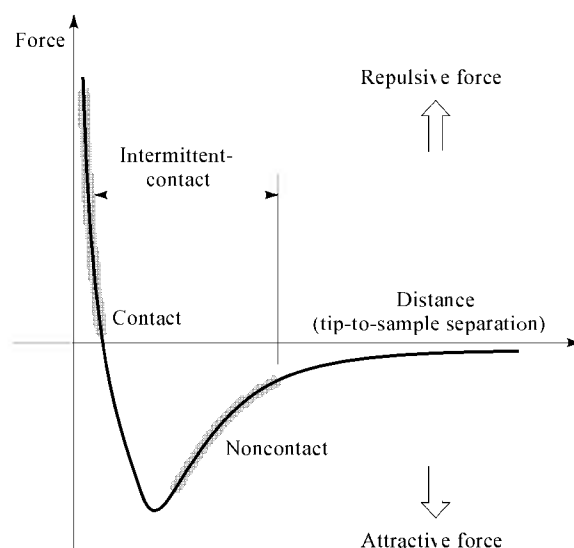


Fig. 6. Van der Waals potential energy-distance curve showing regions of operation of contact, noncontact, and intermittent contact or tapping-mode afm (18).

A second common mode of operation is tapping-mode or intermittent-contact afm. This approach utilizes a vibrating cantilever that intermittently contacts or "taps" the surface, causing a change in oscillation amplitude of the cantilever in response to surface topography. These changes are monitored using an optical detection scheme similar to that described for contact-mode afm. This mode of operation is useful, because it overcomes the potentially destructive effects of contact afm due to the elimination of lateral forces experienced during dragging of the tip while still retaining sub-nm vertical resolution.

New types of scanning probe microscopies are continually being developed. These tools will continue to be important for imaging of surfaces at atomic-scale resolution.

Analysis of Surface Elemental Composition. A very important class of surface analysis methods derives from the desire to understand what elements reside at the surface or in the near-surface region of a material. The most common techniques used for determination of elemental composition are the electron spectroscopies in which electrons or x-rays are used to stimulate either electron or x-ray emission from the atoms in the surface (or near-surface region) of the sample. These electrons or x-rays are emitted with energies characteristic of the energy levels of the atoms from which they came, and therefore, contain elemental information about the surface. Only the most important electron spectroscopies will be discussed here, although an array of techniques based on either the excitation of surfaces with or the collection of electrons from the surface have been developed for the elucidation of specific information about surfaces and interfaces.

X-ray Photoelectron Spectroscopy. X-ray photoelectron spectroscopy (xps) and Auger electron spectroscopy (aes) are related techniques (19) that are initiated with the same fundamental event, the stimulated ejection of an electron from a surface. The fundamental aspects of these techniques will be discussed separately, but since the instrumental needs required to perform such methods are similar, xps and aes instrumentation will be discussed together.

Xps is based on the photoelectric effect when an incident x-ray causes ejection of an electron from a surface atom. Figure 7 shows a schematic of the process for a hypothetical surface atom. In this process, an incident x-ray photon of energy $h\nu$ impinges on the surface atom causing ejection of an electron, usually from a core electron energy level. This primary photoelectron is detected in xps.

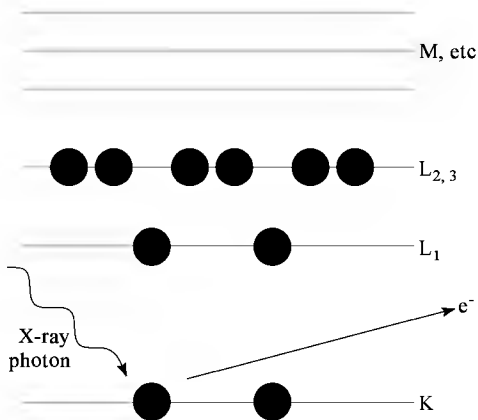


Fig. 7. Schematic of x-ray photoelectron spectroscopy process.

Xps was first successfully implemented for surface analysis by K. Siegbahn in the early 1950s. Much of the development of this approach for studying surfaces was subsequently developed by this same group (20), and Siegbahn was awarded the Nobel Prize in physics in 1981 for these efforts. In order for the primary photoelectron, which is bound to the surface atom with binding energy E_B to be detected in xps, the electron must have sufficient kinetic energy to overcome, in addition to E_B , the overall attractive potential of the spectrometer described by its work function, Φ_{SP} . Thus, the measured kinetic energy of this photoelectron, E_K , is given by

$$E_K = h\nu - E_B - \Phi_{SP} \tag{1}$$

If the energy of the incident x-rays and the spectrometer work function are known, the measured kinetic energy can be used to determine the binding energy E_B from

$$E_B = h\nu - E_K - \Phi_{SP} \tag{2}$$

This E_B is characteristic for each energy level in the element and can be used to determine the element from which the electron originated. Differences in E_B form the basis of elemental sensitivity in xps.

A typical x-ray photoelectron spectrum consists of a plot of the intensity of photoelectrons as a function of electron E_B or E_K . A sample is shown in Figure 8 for Ag (21). In this spectrum, discrete photoelectron responses from the core and valence electron energy levels of the Ag atoms are observed. These electrons are superimposed on a significant background from the Bremsstrahlung radiation inherent in nonmonochromatic x-ray sources (see below) which produces an increasing number of photoelectrons as E_K decreases. Also observed in the spectrum are lines due to x-ray excited Auger electrons.

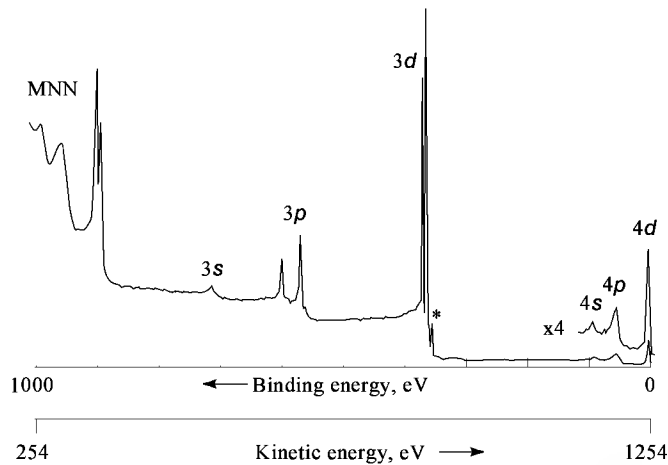


Fig. 8. Xps spectrum of Ag acquired with $Mg K_{\alpha 1,2}$ excitation recorded with constant analyzer energy of 100 eV. (* Ag 3d peak excited with $Mg K_{\alpha 3,4}$ satellite line) (19).

The lines of primary interest in an xps spectrum are those reflecting photoelectrons from core electron energy levels of the surface atoms. These are labeled in Figure 8 for the Ag 3s, 3p, and 3d electrons. The sensitivity of xps toward certain elements, and hence the surface sensitivity attainable for these elements, is dependent upon intrinsic properties of the photoelectron lines observed. The parameter governing the relative intensities of these core level peaks is the photoionization cross-section, ς . This parameter describes the relative efficiency of the photoionization process for each core electron as a function of element atomic number. Obviously, the photoionization efficiency is not the same for electrons from the same core level of all elements. This difference results in variable surface sensitivity for elements even though the same core level electrons may be monitored.

The values commonly used for ς are those calculated by Scofield (22) or Gryzinsky (23). For example, Figure 9 shows Scofield's calculated ς values relative to C 1s electrons for Al K_{α} x-ray radiation. Periodic Table and core level trends are readily apparent in these values. For a given core level, ς generally increases with atomic number (Z).

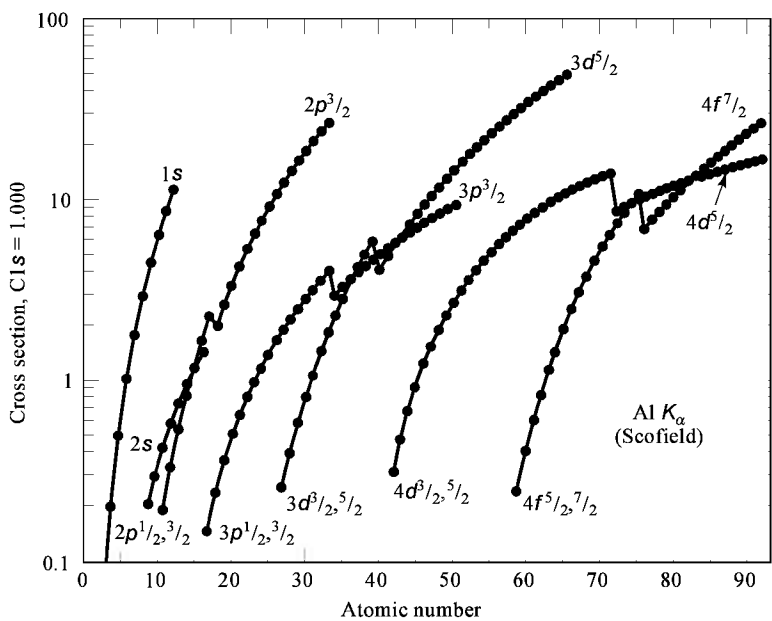


Fig. 9. Scofield's x-ray photoionization cross-section relative to that for C 1s electrons as a function of atomic number (19).

Figure 10 shows a comparison of Scofield's calculated values with experimentally measured values (24) which, in addition to ζ , are dependent on spectrometer transmission function. The overall agreement between the calculated and experimental values is quite good. The far right y-axis in Figure 10 indicates the experimentally accessible surface sensitivities in monolayers (ML) as a function of atomic number. For most elements, sensitivities on the order of 1% of a ML are achievable.

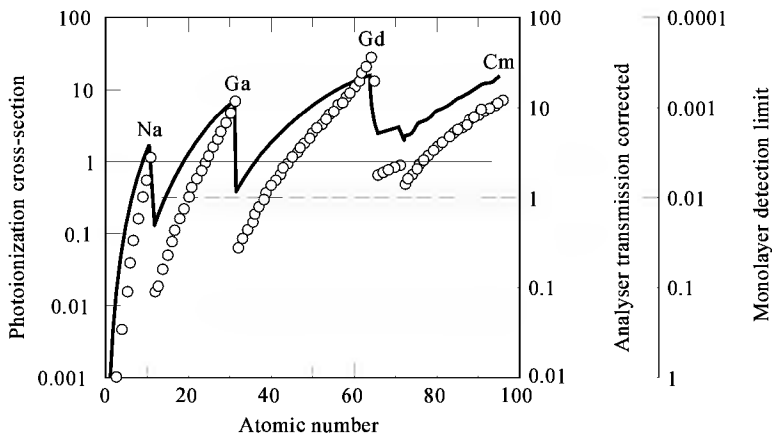


Fig. 10. Comparison of Scofield's calculated x-ray photoionization cross-sections relative to that for F 1s electrons and experimental values (22) as a function of atomic number (3).

Xps is a surface sensitive technique as opposed to a bulk technique because electrons cannot travel very far in solids without undergoing energy loss. Thus, even though the incident x-rays penetrate the sample up to relatively large depths, the depth from which the electron information is obtained is limited by the "escape depth" of the photoemitted electrons. This surface sensitivity of xps is quantitatively defined by the inelastic mean free path parameter which is given the symbol λ . This parameter is defined to be the distance an electron travels before engaging in an interaction in which it experiences an energy loss.

λ is a complex quantity that one would expect to depend on the nature of the solid sample. However, using an extensive database, Seah and Dench (25) compiled a "universal curve" of λ as a function of electron energy in a benchmark paper. This universal curve is shown in Figure 11. The data points in Figure 11 are experimentally measured values and the solid line is the "best fit" line for the experimental data. For the range of electron energies typically encountered in xps (ca 10–1000 eV), λ ranges only from ca 0.5–2 nm. Thus, this represents the range of depths over which the xps experiment typically samples. Seah and Dench derived the following semiempirical relationships for electrons of energies between 1 and 6,000 eV:

For elements

$$\lambda = \frac{538}{E^2} + 0.41a^{0.5}E^{0.5} \tag{3}$$

For inorganic materials

$$\lambda = \frac{2170}{E^2} + 0.72a^{0.5}E^{0.5} \tag{4}$$

In these equations, a , the monolayer thickness of the matrix material comes from

$$AW = 1000\rho n N_A a^3 \tag{5}$$

where AW is the mean atomic weight of the matrix elements, ρ is the bulk material density in kg m⁻³, n is the number of atoms in the molecule and N_A is Avogadro's number.

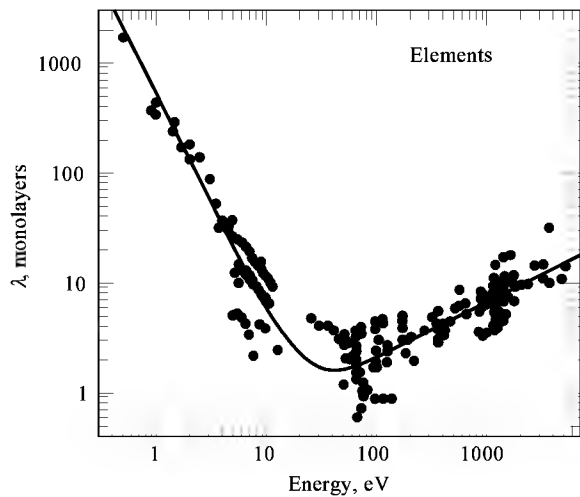


Fig. 11. "Universal curve" of inelastic mean free path, λ , as a function of electron kinetic energy. Solid line is universal curve, points are experimental data compiled for pure elements (19).

With the λ parameter in hand, the Beer-Lambert expression describes how the photoelectron intensity will decrease as a function of depth into the sample:

$$I = I_0 \exp\left(-\frac{z}{\lambda \cos \theta}\right) \tag{6}$$

where I is the intensity at a distance z above the origin of the electron, I_0 is the intensity at $z = 0$, and θ is the collection angle (or "take-off" angle) with respect to the surface normal. One should note that the collection (or "take-off") angle θ can be further exploited as a means of altering surface sensitivity. As θ increases, the depth from which photoelectrons are observed decreases, and therefore, the surface sensitivity of the analysis increases. Thus, angle-resolved xps (arxps) provides a crude method for depth profiling the surface region of a solid sample through control of analyzer collection angle for photoelectrons. The more traditional way of depth profiling using ion sputtering is discussed further below in the section on Auger electron spectroscopy.

One other very important attribute of photoemitted electrons is the dependence of their kinetic energy on chemical environment of the atom from which they originate. This feature of the photoemission process is called the chemical shift of E_B and is the basis for chemical information about the sample. In fact, this feature of the xps experiment, first observed by Siegbahn in 1958 for a copper oxide overlayer on a copper surface, led to his original nomenclature for this technique of electron spectroscopy for chemical analysis or esca.

Chemical shifts can be observed for atoms in both organic and inorganic materials. Generally, E_B increases as the oxidation state of the atom increases. This dependence can be rationalized on the basis of simple electrostatic considerations in terms of an increased attraction for the electron by the atom making it that much harder to remove. These chemical shifts can range in magnitude from <1 eV to >10 eV.

Figure 12 shows xps spectra in the Ti $2p$ region for Ti and TiO_2 (19). Two peaks are observed for the Ti $2p$ electrons from spin-orbit coupling producing the $2p_{3/2}$ and $2p_{1/2}$ peaks at 453.8 and 460.0 eV, respectively. These $2p_{3/2}$ and $2p_{1/2}$ peaks shift ca 5 eV to higher binding energies of 458.5 and 464.2 eV, respectively, for electrons from Ti^{4+} atoms in the TiO_2 .

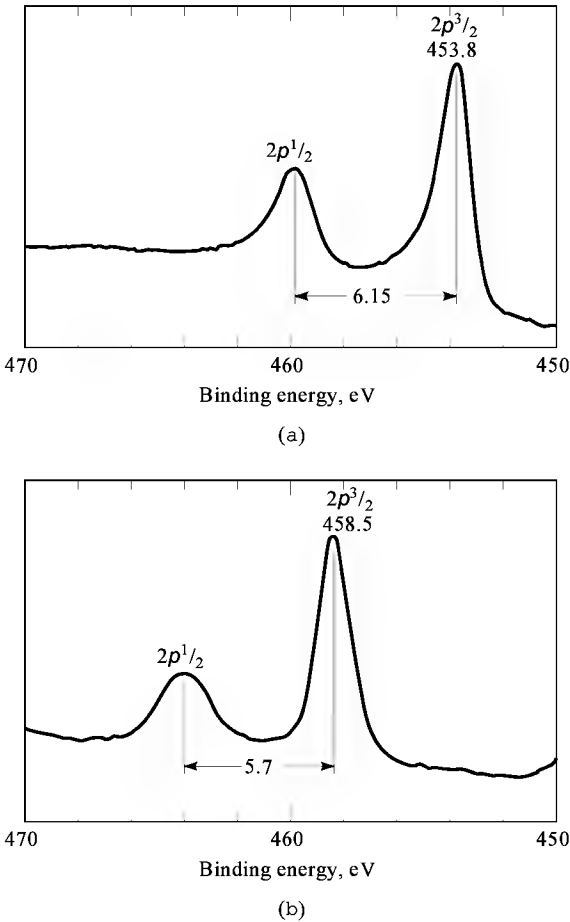


Fig. 12. Xps spectra for Ti $2p$ electrons in (a) Ti, and (b) TiO_2 (21).

Additional examples of the effect of atom oxidation state on peak position are shown in Figure 13 for two polymer systems (26). The spectrum in Figure 13a shows the C 1s region for poly(methyl methacrylate). Distinct xps peaks are observed for the four chemically distinct C environments in this polymer. These are labeled in Figure 13a. The C atoms labeled as 1 exhibit a binding energy of ca 285 eV for C atoms in a hydrocarbon-like environment. The C atom labeled as 4 shows a 5 eV shift to higher binding energies from C 1 and is assigned to the C atom in the highest oxidation state, that of the carbonyl group.

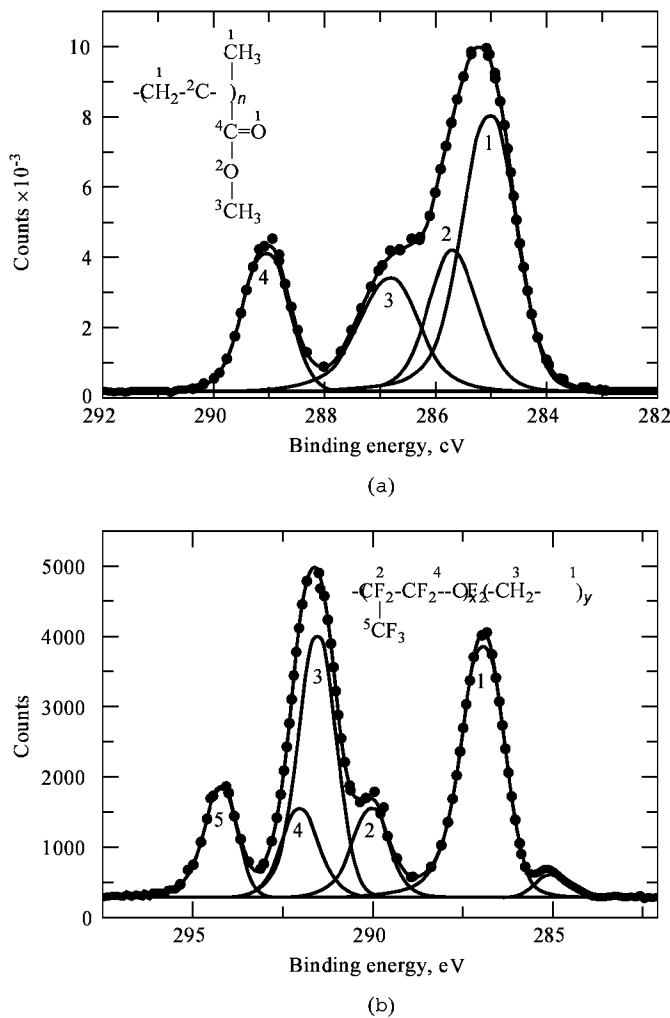


Fig. 13. (a) Xps spectrum for C 1s electrons in poly(methyl methacrylate). (b) Xps spectrum for C 1s electrons in Viton A, a fluoropolymer (26).

A similar example is shown in Figure 13b for the fluorinated polymer Viton A (26). In this case, five chemically distinct C 1s environments exist and are seen discretely by xps. These are labeled in Figure 13b. The C 1s peak at 287 eV is due to C atoms bonded to hydrogen atoms. The four C atoms bonded to very electronegative F atoms exhibit C 1s peaks at ca 290, 291, 292, and 294 eV, higher in energy on the basis of the electronegative F atoms withdrawing electron density leaving the C atom slightly positively charged. Thus, as these examples demonstrate, xps can be used to infer chemical information about surfaces and interfaces through the chemical shift parameter.

Xps spectra also bear a relationship between photoelectron intensity and number of surface atoms sampled (19,27). Quantitation of these data can be achieved with a precision to within ca 20%. For a homogeneous sample analyzed in a fixed geometry, the relationship between xps intensity and number of atoms is given by

$$I = I_0 T D \sigma L \int_0^\infty N \exp\left(\frac{z}{\lambda \cos \theta}\right) dz \tag{7}$$

where I is the xps intensity per unit area illuminated by the x-ray source, I_0 is the incident x-ray flux, T is the transmission efficiency of the analyzer at the kinetic energy of the electron, σ is the photoionization cross-section for the appropriate incident x-ray energy, L is a term that takes into account the effects of angle of incident x-rays and angle of collection of photoelectrons, N is the number of atoms sampled, z is the depth from which the photoelectrons originate, and λ is the inelastic mean free path for photoelectrons at the given kinetic energy.

It is extremely difficult to know values for all of these parameters precisely. Therefore, absolute quantitation is almost never attempted. The determination of relative atomic ratios is an inherently more tractable approach, however. This method is best illustrated by consideration of a binary material composed exclusively of atoms A and B that is perfectly homogeneous up to the surface. In this case, independent equations can be developed relating the number of atoms sampled to the xps intensity for each atom as follows:

$$N_A = \frac{I_A}{I_0 T_A D_A \sigma_A L_A \lambda_A \cos \theta}$$
$$N_B = \frac{I_B}{I_0 T_B D_B \sigma_B L_B \lambda_B \cos \theta} \tag{8}$$

where the terms have the identities described above. For a given experiment in which both xps lines are measured, I_0 and $\cos \theta$ are identical in each equation. Furthermore, if the xps lines of atom A and atom B are close in energy, then $T_A \approx T_B$, $D_A \approx D_B$, $L_A \approx L_B$, and $\lambda_A \approx \lambda_B$. In this case, the ratio N_A/N_B is given by

$$\left(\frac{N_A}{N_B}\right) = \frac{I_A \sigma_B}{I_B \sigma_A}$$

(9)

In many cases, this binary material will not be homogeneous all the way up to the surface, because it is covered with a thin overlayer of contamination. Therefore, for most real samples, the photoelectrons of interest from atoms A and B are coming from a depth equal to the thickness of the overlayer, *d*. In this case,

$$\left(\frac{N_A}{N_B}\right) = \frac{I_A \sigma_B \lambda_B \exp\left(\frac{d}{\lambda_B \cos \theta}\right)}{I_B \sigma_A \lambda_A \exp\left(\frac{d}{\lambda_A \cos \theta}\right)}$$

(10)

The above treatment is predicated on the assumption that the kinetic energies of the photoelectrons from atoms A and B are close in energy. In the event that this assumption does not hold, then all of the instrumental parameters do not cancel for these equations, and the situation is more complex. An alternative strategy in this case is to compare the spectrum of the unknown material with a spectrum acquired under identical conditions of a pure standard reference material containing A and B that is close in suspected composition to the unknown. In this case,

$$\left(\frac{N_A}{N_B}\right)_{UNK} = \left(\frac{I_A}{I_B}\right)_{UNK} \frac{T_B D_B \sigma_B L_B \lambda_B}{T_A D_A \sigma_A L_A \lambda_A}$$

(11)

and

$$\left(\frac{N_A}{N_B}\right)_{STD} = \left(\frac{I_A}{I_B}\right)_{STD} \frac{T_B D_B \sigma_B L_B \lambda_B}{T_A D_A \sigma_A L_A \lambda_A}$$

(12)

Therefore, the relative atomic ratio of the unknown material can be determined from

$$\left(\frac{N_A}{N_B}\right)_{UNK} = \frac{(I_A / I_B)_{UNK} (N_A / N_B)_{STD}}{(I_A / I_B)_{STD}}$$

(13)

This approach is the most useful for routine quantitative xps analysis.

Auger Electron Spectroscopy. Auger electron spectroscopy (aes) is also based on an electron ejection process like xps, but the electrons that are monitored in aes are secondary electrons (19). These secondary or Auger electrons arise from a process shown schematically in Figure 14. The process occurs after primary electron emission such that a core level hole exists. Incident electrons are conventionally used in aes to stimulate primary electron emission, although incident x-rays can also be used as in xps. The presence of this core level electron hole results in electron relaxation from a valence level to fill the core level hole. When this relaxation occurs, the excess energy that is released stimulates ejection of a secondary or Auger electron from another valence electron energy level. The final state of the atom after Auger electron emission is thus doubly ionized; however, because these vacancies are in valence electron levels further away from the nucleus, this final doubly ionized state is more stable than the initial singly ionized state. The energy of this Auger electron is *independent* of the energy of the incident electron beam and depends only on the energies of the core and valence levels involved. The kinetic energy of the Auger electron is given by the difference in energy between the core level and the valence level from which the electron comes that fills the core level vacancy minus the energy that it takes to remove the Auger electron from the singly ionized atom and the spectrometer work function. Thus, for a KLL Auger electron

$$E_{KL(1)L(2,3)} = E_K - E_{L(1)} - E_{L(2,3)}^*$$

(14)

where $E_{KL(1)L(2,3)}$ is the kinetic energy of the Auger electron, E_K is the energy of the *K* core level, $E_{L(1)}$ is the energy of the valence level from which the electron comes to fill the *K* level core vacancy, and $E_{L(2,3)}^*$ is the binding energy of the $L_{(2,3)}$ level *in the presence of a hole in the $L_{(1)}$ level*. Auger electrons typically range in energy from ca 10 to 2000 eV.

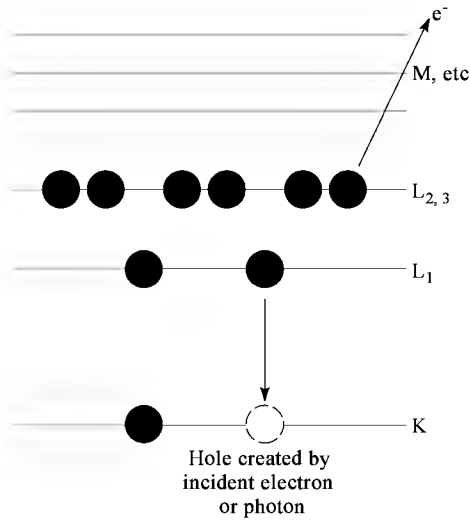


Fig. 14. Schematic of the Auger electron emission process induced by creation of a K level electron hole.

Any three subshells within an atom can be involved in the Auger process as long as the final state is significantly more stable than the initial state. The Auger transition is identified by a three-letter label in which the first letter indicates the core level in which the initial electron vacancy resides, the second letter represents the valence level from which the electron comes that fills the initial vacancy, and the third letter indicates the valence level from which the ejected Auger electron comes. KLL, LMM, and MNN Auger processes are common. Any atom has many possible Auger transitions, but the probability of each of these varies with the energy levels involved. Thus, the sensitivity of aes to different atoms varies. A plot of relative sensitivity factors for different atoms normalized to the Ag 351 eV MNN Auger transition is shown in Figure 15 (28).

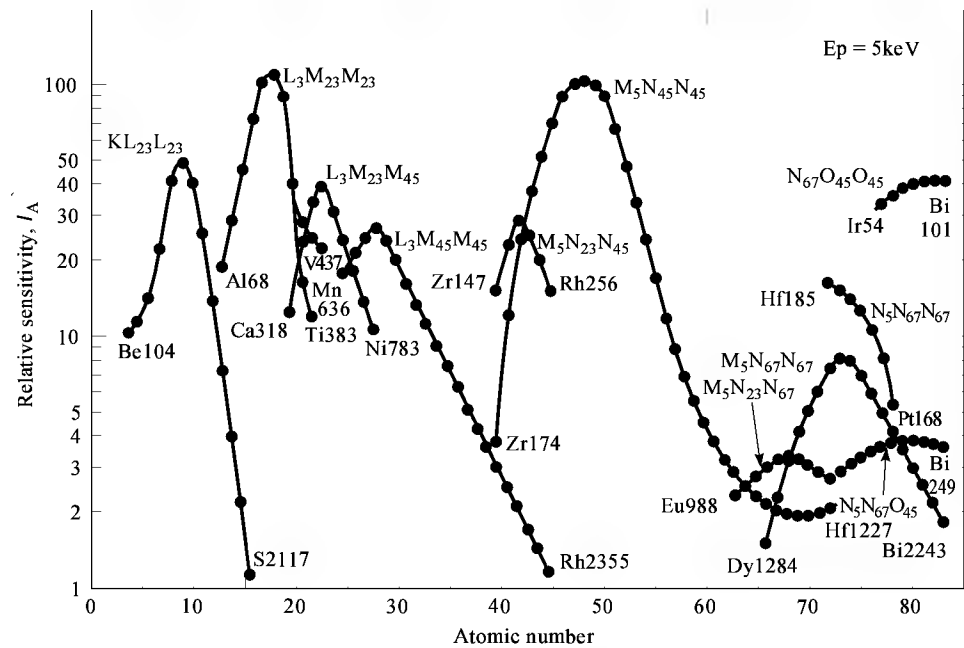


Fig. 15. Auger sensitivity factors relative to the Ag MNN Auger transition as a function of atomic number (19).

An alternative mechanism of excess energy release when electron relaxation occurs is through x-ray fluorescence. In fact, x-ray fluorescence favorably competes with Auger electron emission for atoms with large atomic numbers. Figure 16 shows a plot of the relative yields of these two processes as a function of atomic number for atoms with initial K level holes. The cross-over point between the two processes generally occurs at an atomic number of 30. Thus, aes has much greater sensitivity to low Z elements than x-ray fluorescence.

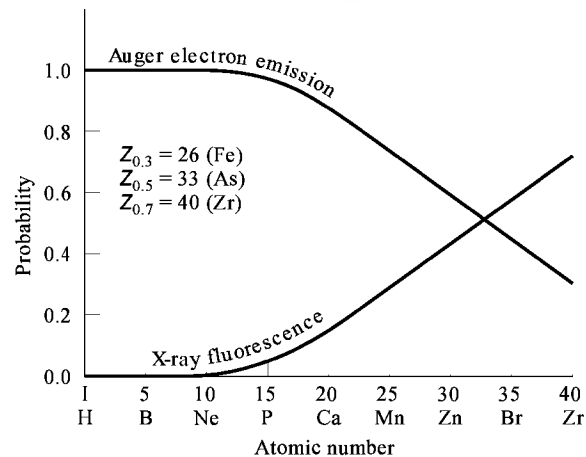


Fig. 16. Relative probabilities of Auger electron emission and x-ray fluorescence for initial K level electron hole as a function of atomic number (19).

Auger spectra are plots of the intensity of Auger electrons as a function of kinetic energy. Figure 17a shows a plot of direct electron intensity as a function of kinetic energy for a copper surface covered with contaminants (19). The Auger electrons of interest are superimposed on a very large background of elastic and inelastically scattered primary electrons used to excite the Auger process. At the lowest kinetic energies, the background increases sharply in an electron "cascade" of inelastically scattered electrons. The intensity of the Auger electrons in relation to this electron cascade is so small as to almost be invisible on this large electron background intensity. Thus, Auger spectra are typically plotted as derivative spectra in which the large background contributions are minimized. An example of such a derivative spectrum is shown in Figure 17b. The Auger peaks are clearly much more visible in this derivative spectrum than they are in the direct intensity spectrum.

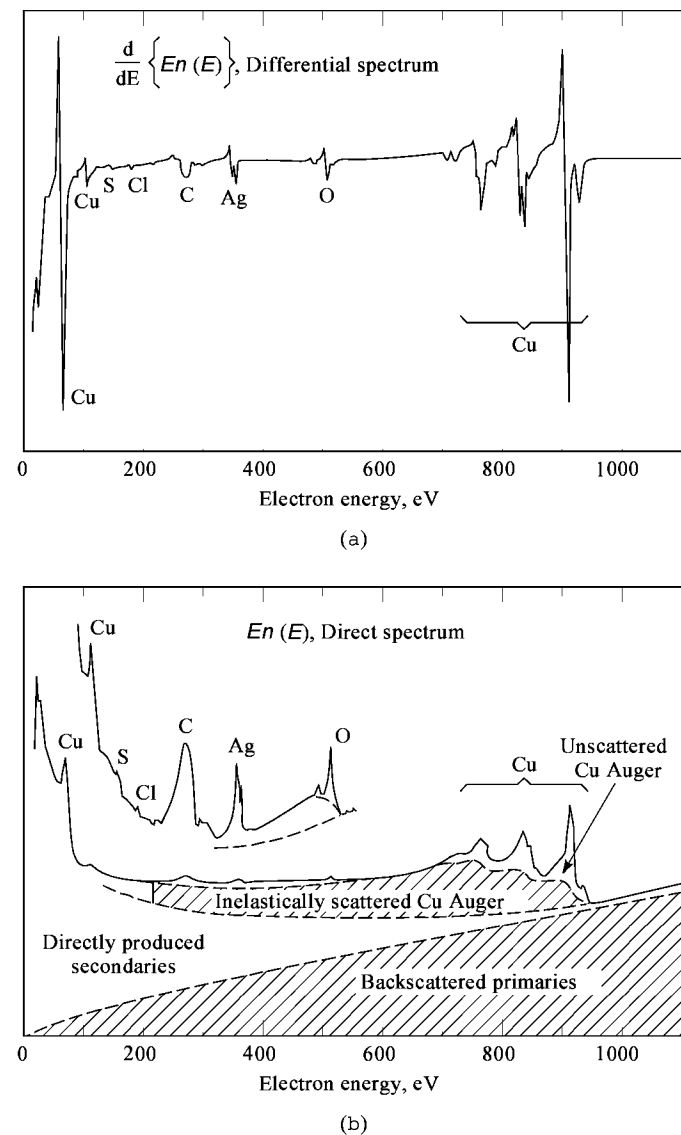


Fig. 17. (a) Auger electron spectrum presented in derivative mode for contaminated Cu surface. (b) Auger electron spectrum presented in direct intensity mode for same sample. Contributions to background intensities indicated (19).

Quantitation of Auger spectra follows the same lines of reasoning as discussed above for quantitation of xps data (19,27). Thus, the current due to Auger electrons at a particular kinetic energy, I , is given by

$$I = I_0 T D \sigma B \int_0^\infty N \exp\left(-\frac{z}{\lambda \cos\theta}\right) dz \tag{15}$$

where I_0 is the primary electron beam current, ζ is the Auger cross-section, T and D are instrumental efficiencies as defined as above, N is the atom number density, λ is the inelastic mean free path, and B is the back-scattering factor. This latter parameter takes on values greater than one and corrects for the creation of Auger electrons by backscattered primary electrons.

Equation 15 is not used directly for quantitation of aes data due to the same limitations discussed above for xps. Particularly troubling for aes is the inability to determine B for a given matrix. Thus, the analyst is left with comparing aes data from an unknown with those of known materials in an attempt to estimate relative atom ratios. This proceeds along the same lines as for xps and allows quantitation of aes data to a precision of typically better than ca 20%.

Aes is often combined with surface ion sputtering using a high energy ion beam to achieve profiling of the elemental composition as a function of depth into the solid (3,19). This approach is typically implemented using a beam of inert gas ions such as Ar^+ accelerated to 200–5000 eV to sequentially sputter away the surface layer by layer. Simultaneous with this sputtering process, aes can be used to probe the elemental composition in the sputtering crater that is created.

An example of how such depth profiling information is presented is shown in Figure 18 for a hypothetical system composed of a substrate of element B covered by a thin film of material containing element A. This figure shows plots of aes intensity as a function of sputtering time for the ideal response and the real response expected from this system. For a given sputtering rate, the sputtering time is related to depth into the sample. If depth profiling of a chemical system with a sharp interface were ideal, and hence sputtering of the material occurred perfectly uniformly, the intensity-sputtering time profiles would mimic the physical sharpness of the interface between A and B. In reality, this does not happen due to a variety of experimental limitations. First, the sputtering process is physically destructive and can lead to significant sputtering-induced roughness which blurs the interface, since parts of B can be exposed before all of A is sputtered away. In addition, the sputtering process can cause implantation of atoms of A into the B substrate, such that their signals are still observed beyond the depth of the original interface. The situation is further complicated by the fact that materials have different rates of sputtering. This aspect of the process can lead to further loss of the real interfacial boundary in the data. All of these features lead to the broadening of the interface in the real depth profiles shown in Figure 18.

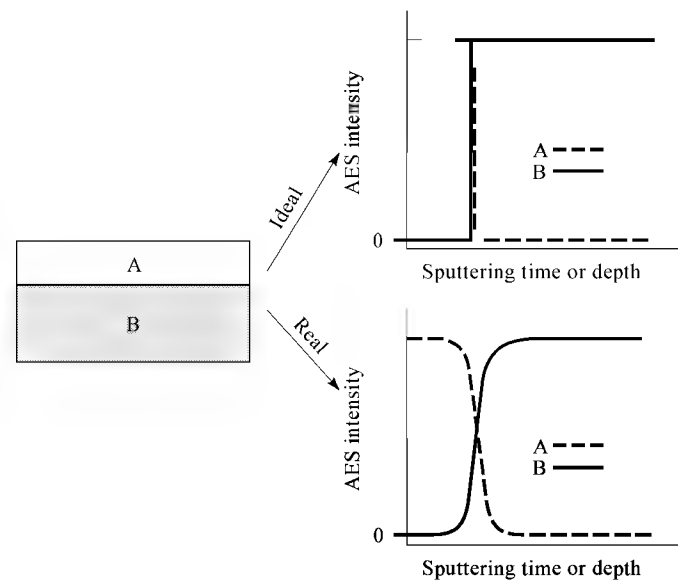


Fig. 18. Ideal and real sputter-depth profiles for thin film of A and substrate B.

Several other aspects of depth profiling by a combination of ion sputtering and aes are noteworthy. It is important that the sputtering ions be chemically nonreactive such that the surface chemistry does not change during the sputtering process. The analyst must also remain cognizant of the possibility of the sputtering process altering the Auger electron emission. As noted above, sputtering is an abrasive process which can lead to the undesirable production of secondary electrons, some of which may be Auger electrons. This problem can be overcome through the use of a much smaller ion beam current density for sputtering than the electron beam current density used to excite aes. The disadvantage of this approach is slower sputtering. An additional consideration when depth profiling by simultaneous sputtering aes is where within the sputtering crater the aes electron beam is positioned. Due to the inherent current density profile of the ion beam, the sputtering crater that is produced is bowl-shaped. Ideally, therefore, sampling in the center bottom of the crater for aes produces information over the shortest depth range. If the sides of the crater are sampled, a much greater range of depths contribute to the aes intensities, further blurring the interface.

Finally, it is difficult to calibrate the depth scale in a depth profile. This situation is made more complicated by different sputtering rates of materials. Despite these shortcomings, depth profiling by simultaneous ion sputtering aes is commonly employed, because it is one of the few techniques that can provide information about buried interfaces, albeit in a destructive manner.

Xps and Aes Instrumentation. The instrumentation required to perform xps and aes analyses is generally sophisticated and expensive (19). The need for UHV conditions in order to retain surface cleanliness for a tractable period of time was mentioned above. Beyond this requirement (and the hardware that accompanies it), the most important components of an electron spectrometer system are the source, the electron energy analyzer, and the electron detector. These will be discussed in turn below.

Xps requires a source that can provide a single x-ray line reasonably narrow in energy. The absolute energy requirement for this x-ray line is that it must be energetic enough to generate photoelectrons from core levels of a majority of the elements with reasonable resolution. Of the many possible x-ray sources, those which best meet this requirement are the Al $K_{\alpha 1,2}$ x-ray line at 1486.6 eV and the Mg $K_{\alpha 1,2}$ x-ray line at 1253.6 eV. These lines have full-widths-at-half-maxima (FWHM) that are ca 0.85 and 0.7 eV, respectively.

X-rays are produced from these materials by electron bombardment using electrons whose kinetic energies are higher than the x-ray energies produced. Once produced, the x-rays leave the source through an x-ray-transparent window typically made of Al. To avoid melting of the x-ray producing material, a thin film (ca 10 μm) of Al or Mg is coated onto a cooled Cu block acting as the anode. Electrons are generated at the cathode. Figure 19 shows how this is commonly done in a dual-anode arrangement in which both x-ray sources are contained within the same housing. In this case, the source can be switched in a few seconds between the Al K_{α} and Mg K_{α} lines.

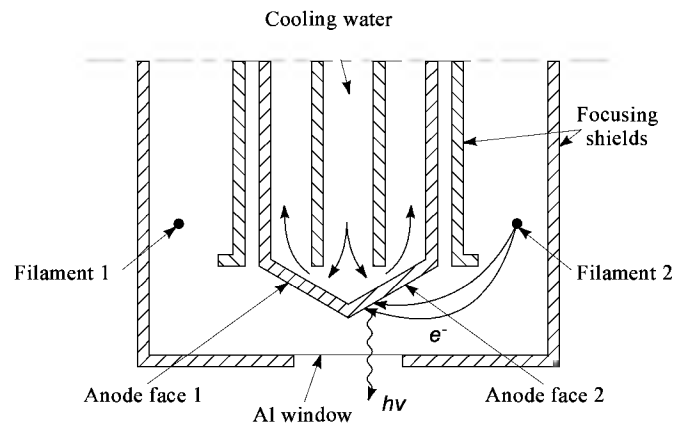


Fig. 19. Schematic of dual anode (typically Al and Mg) x-ray source. X-rays produced by electron bombardment of anode face 2 indicated (19).

The routine dual-anode x-ray source just described generates nonmonochromatized x-rays in a relatively large spot size (ca 1 cm in diameter). In fact, the Al $K_{\alpha 1,2}$ and Mg $K_{\alpha 1,2}$ lines are unresolved doublets resulting from the $2p_{3/2,1/2} \rightarrow 1s$ transitions in these elements. In addition to these lines, a small fraction of the x-rays emitted are in the form of either a continuous background emission called Bremsstrahlung or satellite x-ray lines at other energies (eg, Al $K_{\alpha 3,4}$). The satellite lines result in the corresponding satellite photoelectron lines in the xps spectrum. The Bremsstrahlung x-radiation is more problematic, because it occurs over a continuous band of energies. This Bremsstrahlung radiation is produced when the primary electrons are decelerated from their initial energy in the vicinity of atomic nuclei. Since this deceleration occurs in a nonquantized manner, a continuum of electron energies results in a continuum of x-ray energies. These energies are bounded on the upper end by complete deceleration of the primary electron to zero kinetic energy in a single collision. This upper energy is given by the Duane-Hunt law (29)

$$h\nu_0 = \frac{hc}{\lambda_0} = Ve$$

(16)

where λ_0 is the limiting wavelength in E corresponding to this cut-off energy V is the accelerating voltage of the primary electron, and e is the charge on the electron. Figure 20 shows the output of an Al target under electron bombardment with 15 keV electrons on both a linear and logarithmic scale. The Bremsstrahlung background, although clearly a small fraction of the K_α line, becomes quite noticeable when shown on a logarithmic scale.

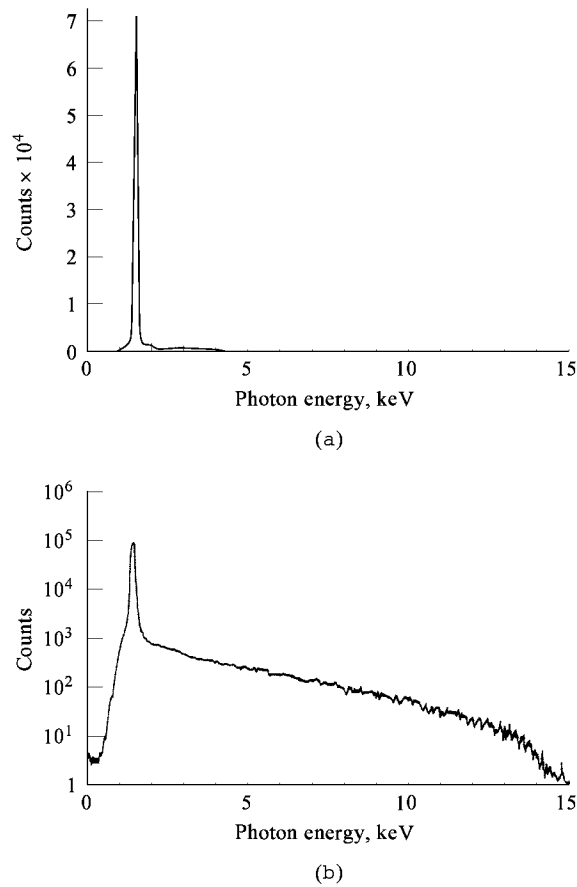


Fig. 20. Primary K_α x-ray line and Bremsstrahlung background excited by bombardment with 15 keV electrons. (a) Linear scale plot. (b) Logarithmic scale plot (19).

In order to achieve higher resolution, the absence of satellite lines, and the effects of the Bremsstrahlung background in the spectrum, the use of monochromatized x-rays is necessary. X-ray monochromators, commercially available for such purposes, are based on Bragg diffraction typically using a toroidal quartz crystal on a Rowland circle. Such sources, while producing narrower x-ray lines, have lower fluxes which result in lower sensitivity. Thus, the analyst must decide if the trade-off between sensitivity and energy resolution is needed for the intended application. More recently, advances in x-ray focusing optics have resulted in the commercial availability of "small spot" xps systems in which the spot size of the x-ray beam can be focused down to ca 30 μm in diameter. These systems, although affording considerably lessened sensitivity, provide greatly enhanced lateral surface resolution.

The source required for aes is an electron gun similar to that described above for electron microscopy. The most common electron source is thermionic in nature with a W filament which is heated to cause electrons to overcome its work function. The electron flux in these sources is generally proportional to the square of the temperature. Thermionic electron guns are routinely used, because they are robust and reliable. An alternative choice of electron gun is the field emission source which uses a large electric field to overcome the work function barrier. Field emission sources are typically of higher brightness than the thermionic sources, because the electron emission is concentrated to the small area of the field emission tip. Focusing in both of these sources is done by electrostatic lenses. Today's thermionic sources typically produce spot sizes on the order of 0.2–0.5 μm with beam currents of 10^{-8} A at 10 keV. If field emission sources are used, spot sizes down to ca 10–50 nm can be achieved.

The heart of the electron spectrometer system is the electron energy analyzer (19). The electrons to be measured in xps and aes have relatively low kinetic energies; therefore, the analyzers rely on electrostatic focusing in which electron trajectories are manipulated using electric fields. The figure-of-merit for analyzers is energy resolution. Analysts refer to analyzer resolution in one of two ways: absolute energy resolution (ie, the FWHM of an electron spectroscopic peak or ΔE) and relative resolution, R, which is given by

$$R = \frac{\Delta E}{E_K}$$

(17)

where E_K is the kinetic energy at the peak position. Relative resolution is often referred to in terms of the spectrometer resolving power, ρ , which is the inverse of the relative resolution.

In most electron spectroscopic analyses, the kinetic energies of the electrons entering the analyzer are retarded to either a constant energy or by a constant factor. These approaches lead to two modes of operation: the constant analyzer energy (CAE) mode and the constant retard ratio (CRR) mode. In the CAE mode, electrons are retarded to a constant pass energy, E_p , that is fixed over the entire spectrum. This mode results in constant absolute resolution throughout the spectrum and is the most commonly employed mode for xps. This retardation results in an improvement in relative resolution by a factor equal to E_p/E_K compared to the resolution before retardation. This mode is tantamount to providing constant analyzer transmission throughout a spectral scan in xps and makes the spectra easier to quantify, since ΔE is constant. The trade-off, however, is that at low kinetic energies, the S/N of the signal degrades considerably due to the large increase in background electron intensity.

The CRR mode involves retarding the electron kinetic energies to a constant ratio of E_K/E_p where E_p is the energy passed by the analyzer. Thus, the energies are retarded by a constant factor. Spectra acquired in this mode are less easy to quantify, but small peaks at low kinetic energies are readily detected. This mode of operation results in spectra of constant relative resolution throughout. The relative resolution is improved in this mode by a factor of E_p .

These modes of operation are used in conjunction with the two most popular energy analyzers, the cylindrical mirror analyzer (CMA) and the concentric hemispherical analyzer (CHA). The most common form of the CMA used today is the double-pass version diagramed in Figure 21. This device consists of two perfectly coaxial cylinders of radii r_o and r_i . The outer cylinder is held at a potential of $(-V_o)$ and the inner cylinder is held at ground. The

spatial separation of electron energies in this device can be understood by consideration of the behavior in the first stage of the CMA shown in Figure 21. Electrons of kinetic energy E_K emitted from a source (ie, the sample) which lies exactly on the axis of the CMA are collected by the analyzer, usually at an angle, α , of 42° with respect to the analyzer axis. The kinetic energy that is transmitted by the analyzer is given by

$$E_K = \frac{K e V_o}{\ln\left(\frac{r_o}{r_i}\right)} = \frac{1.31 e V_o}{\ln\left(\frac{r_o}{r_i}\right)} \tag{18}$$

where the constant K is 1.31 for $\alpha = 42^\circ$. Obviously, a range of acceptance angles, $\Delta\alpha$, must be used for efficient collection. E_K is systematically varied by sweeping V_o for the scan of a spectrum. For a slit width of w , the energy resolution for the CMA used at this acceptance angle is given by

$$\frac{\Delta E}{E_K} = \left(\frac{0.18 w}{r_i}\right) + 1.39(\Delta\alpha)^3 \tag{19}$$

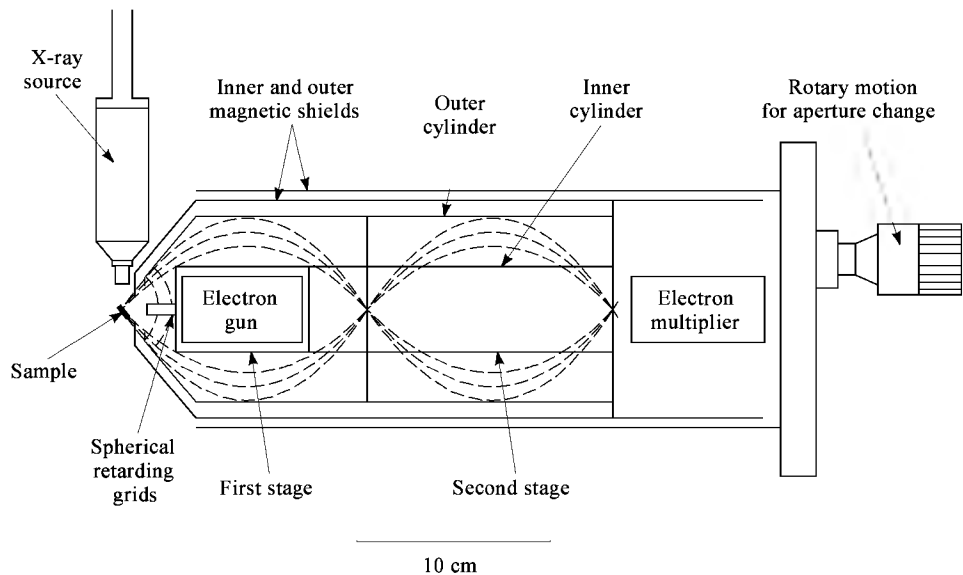


Fig. 21. Double-pass cylindrical mirror analyzer (19).

In xps in which small chemical shifts are employed to ascertain chemical information, energy resolution of the analyzer is an important consideration. Although resolution can be increased by decreasing $\Delta\alpha$, this is done at the expense of sensitivity as dictated by the analyzer luminosity (the product of the analyzer transmission and the acceptance area). Thus, in effectively using the CMA, the analyst is faced with the need to optimize both resolution and luminosity. The second stage of a double-pass CMA becomes critical for achieving these divergent goals. In typical use of a double-pass CMA for xps, the first stage is set to control resolution through $\Delta\alpha$ and the second stage to control luminosity by maximizing transmission of the system to the point at which resolution just begins to be degraded. The optimized situation is most commonly achieved in the CAE mode of operation for xps.

In aes, the resolution is largely independent of the characteristics of the analyzer or source and is dictated by the natural linewidth of the Auger line (usually several eV). Therefore, in using a CMA for aes, the analyst is more concerned with transmission (and hence, sensitivity) than with resolution. In contrast to xps, the optimization of variables is achieved for aes in the CRR mode of operation. The large transmission of the CMA relative to the CHA make it the more desirable analyzer for aes.

The CHA is used for xps today in most commercial spectrometers. The most common version of the CHA is the 180° device shown in Figure 22. It consists of two concentric hemispherical surfaces of radii R_2 and R_1 . These surfaces have a potential difference of ΔV applied between them so that the outer surface is negative and the inner surface is positive. The median equipotential surface R_0 falls between these surfaces; ideally, $R_0 = (R_1 + R_2)/2$. The kinetic energy, E_K , of an electron that passes through the CHA is given by

$$E_K = \frac{e \Delta V}{\left(\frac{R_2}{R_1} - \frac{R_1}{R_2}\right)} \tag{20}$$

For slits placed at R_0 from the center of curvature, the electrons passed by this analyzer follow the equipotential surface described by R_0 . With an acceptance angle $\delta\alpha$ shown in Figure 22 and a slit width w , the energy resolution of the CHA is given by

$$\frac{\Delta E}{E_K} = \left(\frac{w}{2R_0}\right) + \left(\frac{\delta\alpha^2}{4}\right) \tag{21}$$

The value of α is typically chosen to be $\alpha \approx w/2R_0$ to optimize the compromise between sensitivity and resolution. In this case, the resolution becomes

$$\frac{\Delta E}{E_K} = \frac{0.63 w}{R_0} \tag{22}$$

Thus, for a set R_0 , an inverse relationship between pass energy, E_K , and slit width, w , exists. Retardation with a CHA can either be in the CAE or CRR modes. This retardation is accomplished by planar retarding grids placed across the entrance aperture of the CHA.

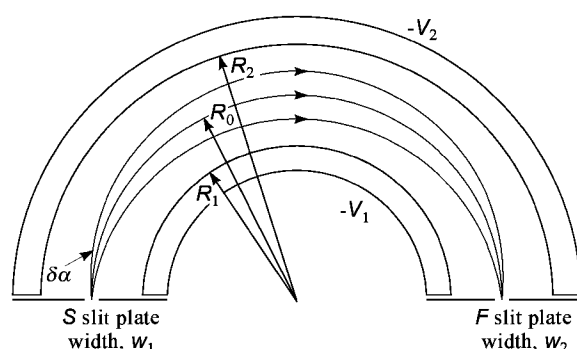


Fig. 22. Concentric hemispherical analyzer (19).

The most common detector for electron spectroscopy is the channel electron multiplier. The raw currents encountered for surface-ejected electrons in a typical xps or aes experiment are on the order of 10^{-14} to 10^{-16} A, far too low to be measured directly. Channel electron multipliers consist of a long curved surface tapering to a point coated along the entire inside with a highly resistive material which acts as an electron amplifier. A potential difference of several thousand volts is applied across the device from front to back which accelerates electrons. When electrons emerging from the analyzer impinge on the inside surfaces of the electron multiplier, multiple electrons are ejected. These electrons are accelerated by the potential field down the channel until the next portion of the surface is encountered at which more electrons are ejected due to the increasing kinetic energy of the electrons as they travel down the potential field. Thus, electron amplification on the order of 10^7 to 10^8 is achieved making the currents produced by the electron multiplier easily measurable.

Electron Microprobe Analysis. Electron microprobe analysis (ema) is a technique based on x-ray fluorescence from atoms in the near-surface region of a material stimulated by a focused beam of high energy electrons (7–9,30). Essentially, this method is based on electron-induced x-ray emission as opposed to x-ray-induced x-ray emission, which forms the basis of conventional x-ray fluorescence (xrf) spectroscopy (31). The microprobe form of this x-ray fluorescence spectroscopy was first developed by Castaing in 1951 (32), and today is a mature technique. Primary beam electrons with energies of 10–30 keV are used and sample the material to a depth on the order of 1 μm . X-rays from all elements with the exception of H, He, and Li can be detected.

The ability to perform x-ray analysis is commonly encountered as an added feature of sem instrumentation most often as energy dispersive x-ray analysis (edx) as discussed above (7–9). Ema usually refers to a stand-alone instrument for electron-induced x-ray analysis in either the edx or wavelength dispersive x-ray spectroscopy mode (wds). Ema is used for the determination of elemental constituent identification due to the dependence of x-ray energy on atomic number (Z) of the atom from which the x-ray originates. Ema is also commonly employed for compositional mapping of elements across a surface. Using today's technology, ema can provide elemental maps of both major and minor constituents at detectabilities of a few hundred parts per million in the best cases. This elemental mapping can be accomplished with a lateral resolution of ca 1 μm .

The limit of detection of ema arises from the presence of significant background x-ray intensity, predominantly Bremsstrahlung radiation. The source of this radiation was described above in the section on x-ray sources for xps. Bremsstrahlung radiation is particularly troubling when analyzing high Z elements due to the fact that this radiation competes with the primary electron beam as an excitation source for x-ray fluorescence. The intensity of Bremsstrahlung radiation is function of Z and the difference between incident electron energy and a particular energy of the Bremsstrahlung radiation. Kramers' equation (33) describes this Bremsstrahlung intensity

$$I_{\lambda} = kZ(E_0 - E_{\lambda}) \quad (23)$$

where I_{λ} is the Bremsstrahlung intensity at an energy E_{λ} , k is a constant, and Z is the atomic number of the x-ray producing material. The result is that for high Z materials, over 90% of the x-ray fluorescence excitation arising from K ionization processes comes from the Bremsstrahlung radiation as opposed to primary electron beam excitation.

X-rays are collected and analyzed in ema in one of two ways. In wds, x-rays are dispersed by Bragg diffraction at a crystal and refocused onto a detector sitting on a Rowland circle. This arrangement is similar to the production of monochromatized x-rays for xps described above. In the other approach, edx, x-rays are all collected at the same time in a detector whose output scales with the energy of the x-ray (and hence, Z of the material which produces the x-ray.) Detectors used for ema today are almost exclusively Li-drifted Si solid-state detectors.

Wds provides higher energy resolution than edx, but at the expense of spectrometer transmission. However, since in wds only a single energy increment is measured at one time, and hence the background only at that one energy increment is measured and not the entire background distribution, the signal-to-noise ratio for wds is usually greater than edx despite the advantage of edx in terms of spectrometer transmission. The trade-off for this better sensitivity for wds is oftentimes sample stability. Prolonged electron beam exposure times are needed to overcome the limited transmission of a wds spectrometer, and samples can thermally degrade or become altered by thermally induced atom migration within the sample.

Ema data can be quantitated to provide elemental concentrations, but several corrections are necessary to account for matrix effects adequately. One well-known method for matrix correction is the zaf method (7,31). This approach is based on calculated corrections for major matrix-dependent effects which alter the intensity of x-rays observed at a particular energy after being emitted from the corresponding atoms. The zaf method corrects for differences between elements in electron stopping power and backscattering (the z correction), self-absorption of x-rays by the matrix (the a correction), and the excitation of x-rays from one element by x-rays emitted from a different element, or in other words, secondary fluorescence (the f correction). Once these corrections are made, quantitative ema can be used to determine concentrations of elements in the near-surface region (ca 1 μm) of samples of interest.

Analysis of Surface Molecular Composition. Information about the molecular composition of the surface or interface may also be of interest. A variety of methods for elucidating the nature of the molecules that exist on a surface or within an interface exist. Techniques based on vibrational spectroscopy of molecules are the most common and include the electron-based method of high resolution electron energy loss spectroscopy (hreels), and the optical methods of ftr and Raman spectroscopy. These tools are tremendously powerful methods of analysis because not only does a molecule possess vibrational modes which are signatures of that molecule, but the energies of molecular vibrations are extremely sensitive to the chemical environment in which a molecule is found. Thus, these methods directly provide information about the chemistry of the surface or interface through the vibrations of molecules contained on the surface or within the interface.

The most common and readily accessible methods are those based on ftr spectroscopy. Several of the more common of these are discussed here.

Transmission Fourier Transform Infrared Spectroscopy. The most straightforward method for the acquisition of ir spectra of surface layers is standard transmission spectroscopy (35,36). This approach can only be used for samples which are partially ir transparent or which can be diluted with an ir transparent medium such as KBr and pressed into a transmissive pellet. The extent to which the ir spectral region (typically ca 600–4000 cm^{-1}) is available for study depends on the ir absorption characteristics of the solid support material. Transmission ftr spectroscopy is most often used to study surface species on metal oxides. These solids leave reasonably large spectral windows within which the spectral behavior of the surface species can be viewed.

Pellets of such oxides are usually prepared by mixing 5–10 wt % of the oxide with KBr and pressing a pellet that is several mm thick. Although the preparation of such systems is not difficult, often the pellet must be pressed for long periods of time to allow the KBr to fuse around the oxide particulates. Once prepared, ir spectra of these pellets can be acquired in standard KBr pellet holders.

The advantage of this approach for characterization of surface species is the ease of sample preparation and straightforward data interpretation. Spectra of such samples resemble normal transmission ir spectra and can be quantitated using conventional methods. In addition, this approach is reasonably sensitive with surface coverages down to ca 10⁰ % of a monolayer easily observed on oxides such as silica and alumina. One important point to note is the spectral interference that water adsorbed on an oxide surface can produce. In some cases, the spectrum from surface-confined water can swamp out the spectrum of other surface species of interest. In these cases, the transmission signal will not be strong enough to be of much use. This approach also does not work well for solids which are inherently strong ir absorbers for similar reasons. In these cases, approaches other than transmission ftr spectroscopy must be used to acquire surface vibrational spectra.

Diffuse Reflectance Infrared Fourier Transform Spectroscopy. An alternative approach to the acquisition of surface ir spectra for particulate or powdered samples is diffuse reflectance infrared Fourier transform spectroscopy (drifts) (35–37). This technique was originally developed for the ir characterization of powdered organic and inorganic samples. Its use was later extended to the study of surface species and interfacial films on particulate samples that were difficult to study by other vibrational spectroscopic methods. The technique is based on the diffuse reflectance of radiation that occurs when it is directed onto a surface with a matte finish or a sample comprised of a powder. This diffuse reflectance is different than specularly reflected radiation in that it penetrates and interacts with a sample before emerging. While the radiation is within the sample, light scattering occurs such that the diffusely reflected light emerges from the sample at all angles. Therefore, efficient collection requires collection of the diffuse reflectance over the entire 2π solid angle above the sample surface. In acquiring diffuse reflectance spectra of bulk materials, the sample is often mixed with a nonabsorbing matrix such as KBr. This matrix can have the effect of increasing the amount of ir radiation that is diffusely reflected. This practice might also be done for the characterization of surface-confined films or molecules, although such samples are more often examined undiluted by such a matrix to increase sensitivity.

The amount of diffusely reflected light cannot be measured directly; instead, it is typically measured relative to a nonabsorbing reference material such as a powdered alkali halide to allow adequate correction for scattering characteristics of the powdered sample. In acquiring drift spectra, both the sample and the reference must be held in sample cells that are sufficiently deep that a further increase in sample cell depth would not change the magnitude of the signal measured. Once this sample cell depth is reached, the sample is said to be sampled at infinite depth. This depth is generally ca 3 mm for most materials. The signal that is measured is the ratio of the diffuse reflectance of the sample to that of a nonabsorbing reference material. This ratio is given the symbol R_∞. Measurement in this way allows quantitation according to Kubelka-Munk theory (38,39) which is defined in terms of the Kubelka-Munk function as

$$f(R_{\infty}) = \frac{(1 - R_{\infty})^2}{2R_{\infty}} - \frac{k}{s}$$

(24)

where R_∞ is defined above, *k* is the absorption coefficient, and *s* is the scattering coefficient. The assumptions made in this derivation are that the specularly reflected component of the signal is small and that *s* is constant. The first assumption is valid for most commercially available drifts attachments. The second assumption relies on the uniformity in particle size between the sample and the reference material, which can also be achieved with care in sample preparation. Importantly, Kubelka-Munk theory predicts a linear relationship between the Kubelka-Munk function and concentration for dilute concentrations.

Several factors affect the bandshapes observed in drifts of bulk materials, and hence the magnitude of the diffuse reflectance response. Particle size is extremely important, since as particle size decreases, spectral bandwidths generally decrease. Therefore, it is desirable to uniformly grind the samples to particle sizes of <50 μm. Sample homogeneity is also important as is the need for dilute concentrations in the nonabsorbing matrix.

For surface systems, the considerations described above change very little. Most surfaces are sampled directly without dilution for maximum sensitivity. For such studies, R_∞ is usually defined as the diffuse reflectance measured from the solid sample containing the surface molecules or surface film ratioed to the diffuse reflectance measured from the same solid material in the absence of the surface molecules or surface film. Using this approach, the Kubelka-Munk function is linear with surface coverage as long as the surface coverages are kept low. At higher surface coverages, the response can become nonlinear. Using this approach for the study of surface systems such as catalysts, extremely low limits of detection can be realized with sensitivities of 1⁰ % of a monolayer or less common. Thus, drifts is a straightforward method to implement, since standard attachments to most commercial ftr spectrometers are readily available. Its ease of implementation makes it a valuable addition to the surface analyst's repertoire for powdered samples.

Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy. Attenuated total reflectance (atr) ftr spectroscopy is based on the principle of total internal reflection (40). Methods based on internal reflection in the uv and visible regions of the spectrum are also common in addition to those in the ir region. The implementation of internal reflection in the ir region of the spectrum provides a means of obtaining ir spectra of surfaces or interfaces, thus providing molecularly-specific vibrational information.

Internal reflection starts by consideration of an interface between two media, a denser transparent medium with refractive index *n*₁, and a rarer medium with a complex refractive index *n*[^]₂ (*n*₂ - *i**k*₂ where *k*₂ is the absorption coefficient of the medium) as shown in Figure 23. If *k*₂ of the rarer medium is zero, then the rarer medium is also transparent. For two transparent phases, total internal reflection occurs when radiation of wavelength λ₁ (= λ *n*₁ where λ is the vacuum wavelength) is incident on the interface at an angle θ_i greater than the critical angle θ_c (= sin⁻¹ [*n*₂ / *n*₁]). Although no net power is transmitted into the rarer medium, an evanescent standing wave is set up parallel to the interface which decays in amplitude exponentially on both sides of the interface. This evanescent wave has an electric field associated with it which samples the near-interface region of the rarer medium and is the basis of spectroscopies based on internal reflection. Thus, this evanescent field is inherently interface-selective because it only samples the interface region and not the bulk.

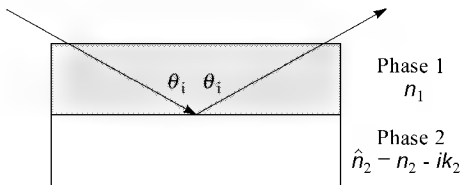


Fig. 23. Total internal reflection occurs at boundary between denser, transparent phase (characterized by *n*₁) and rarer, absorbing phase (characterized by *n*₂) when θ_i is greater than the critical angle, θ_c.

The penetration depth of this evanescent field, *d_p* (defined to be the depth at which the evanescent field decays to 1 / *e* of its original value,) is given by

$$d_p = \frac{\lambda_1}{2\pi \left[\sin^2 \theta_i - (n_2 / n_1)^2 \right]^{1/2}}$$

(25)

The electric field associated with this electromagnetic wave has two notable characteristics which distinguish it from the incident beam and make it useful

in the analysis of surfaces and interfaces. First, the electric field intensity (and hence, light intensity) of the evanescent wave is enhanced relative to the incident beam. Secondly, the evanescent field has amplitude in all three directions in contrast to the incident traveling wave which has amplitude in only one direction.

Attenuated total reflection, on which atr–ftir is based, occurs when the rarer medium is absorbing and is characterized by a complex refractive index $\hat{n}_2 = n_2 + ik_2$ (40). The absorbing characteristics of this medium allow coupling to the evanescent field such that this field is attenuated to an extent dependent on k_2 . The critical angle in the case of attenuated total reflection loses its meaning, but internal reflection still occurs. Thus, if the internally reflected beam is monitored, its intensity will reflect the loss associated with the internal reflection process at the interface with an absorbing medium.

The extent of the interaction between the evanescent field and the absorbing medium is formally described by the effective thickness, d_e , of this medium. This effective thickness is the thickness of the absorbing phase that would have to be passed through by the incident beam in a transmission experiment to give the same energy loss as in the attenuated total reflection experiment. The exact expressions for effective thickness can be very complex. However, for a single attenuated total reflection of an incident beam of radiation of electric field E_i that occurs at an interface between two bulk phases (ie, phase 2 is not a thin film), d_e is given by

$$d_e = \frac{(n_2 - n_1)}{\cos \theta_i} \int_0^\infty E_i^2 dz \tag{26}$$

where $E = E_i \exp(-z/d_p)$. This expression simplifies to

$$d_e = \frac{(n_2 - n_1) E_i^2 d_p}{\cos \theta_i} \tag{27}$$

when k_2 is very small (ie, $\ll 1$). It is important to note in the expression that the strength of coupling of the incident beam with medium 2 decreases (ie, smaller d_e) as θ_i increases. From this expression for effective sampling thickness, one can develop a Beer's law expression for internal reflection based on the relationship between absorbance, A , and reflectance, R (the fraction of light reflected at an interface) as given by

$$A = -\log R = \frac{\alpha_2 n_2 d_p I_0}{2.303 n_1 \cos \theta_i} \tag{28}$$

where $I_0 = E_i^2$ and $\alpha_2 = 4\pi k_2 / \lambda_1$.

The real utility of d_e comes in the analysis of thin films. Consider a substrate of refractive index n_3 supporting a thin film of thickness d and refractive index n_2 in contact with an internal reflection element (the prism) of refractive index n_1 as shown in Figure 24. In this case, d_e depends on the polarization of the incident light beam and is given by

$$d_{e,s} = \frac{4(n_2 - n_1) d \cos \theta_i}{\left(1 - (n_3 - n_1)^2\right)}$$
$$d_{e,p} = \frac{4(n_2 - n_1) d \cos \theta_i \left[\left(1 + (n_3 - n_2)^4 \sin^2 \theta_i - (n_3 - n_1)^2\right) \right]}{\left(1 - (n_3 - n_1)^2\right) \left[\left(1 + (n_3 - n_1)^2 \sin^2 \theta_i - (n_3 - n_1)^2\right) \right]} \tag{29}$$

The film thickness d is what would be sampled in a simple ir transmission experiment. Thus, if d_e is the same as d , then the internal reflection experiment samples the same amount of the thin film as the transmission experiment, and no advantage accrues to the use of internal reflection. If, on the other hand, d_e is greater than d , then the effective sampling of the thin film in an internal reflection experiment is greater than in a transmission measurement, and the internal reflection experiment offers greater sensitivity. Conversely, if d_e is less than d , then a transmission measurement provides the greater sensitivity of the two measurement strategies. It should be noted that, according to the above expressions, optimization of d_e is afforded to a certain extent by alteration of θ_i and n_3 .

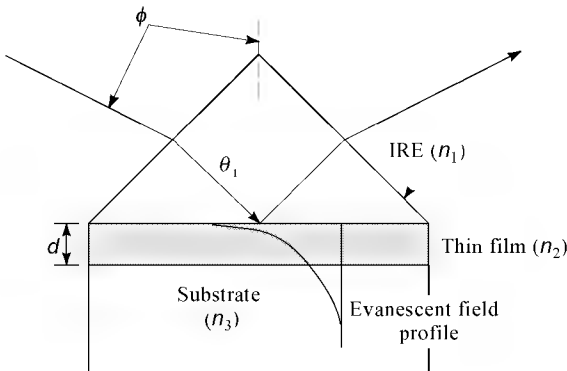


Fig. 24. Attenuated total reflectance of thin film of thickness d and refractive index n_2 on a substrate of refractive index n_3 at the surface of internal reflection element (IRE) of refractive index n_1 . Decay of evanescent field beyond thickness of thin film indicated.

The above discussion was based on the premise of a single internal reflection. An increase in sensitivity can be realized through the use of multiple internal reflections. In the case of multiple reflections, the absorbance A becomes

$$A = -\log R^N = -\log (1 - \alpha d_e)^N \tag{30}$$

for N reflections.

Atr–ftir can be readily performed on most commercial ftir spectrometers through the use of an attachment for atr spectroscopy. These devices provide ir-transparent internal reflection elements that are typically made of Ge, KRS-5, ZnSe, or ZnS. These internal reflection elements are made of materials that are of extremely high purity to avoid losses from absorption by impurities in these devices. Coupling of a thin film or surface sample to one of these reflection elements is accomplished by pressing the sample against the element while acquiring the spectrum.

Infrared Reflection–Absorption Spectroscopy. For adsorbed surface species or thin films on ir reflective surfaces such as metals, an alternative method is infrared reflection–absorption spectroscopy (irras) (41). This technique is based on the external reflection on an infrared beam of light at such surfaces, the characteristics of which are highly polarization dependent. Upon reflection of a polarized beam of light at a surface, phase shifts occur that can be quite significant in magnitude (42,43). Given that the electric field experienced by a molecule at a reflective surface equals the sum of the fields from the incident and reflected beams, the magnitude of this phase shift is critical in determining the resulting light intensity available with which a surface molecule can couple. Upon reflection of an s-polarized ir beam (ie, polarized perpendicular to the plane of incidence, and hence, parallel to the surface), the phase shift is almost 180°. Thus, the resultant electric field at the surface is almost zero due to cancellation of the two electric field vectors. However, for p-polarized light (ie, parallel to the plane of incidence and hence, perpendicular to the surface), the phase shift is much less than 180°. The resultant field at the surface is enhanced due to the summation of the electric fields of the incident and reflected beams. For this situation, the light intensity maximizes at near-grazing angles of incidence. A schematic of these phenomena is shown in Figure 25. These experiments can be performed using standard attachments for ftr spectrometers that can be purchased in either fixed angle or variable angle geometries.

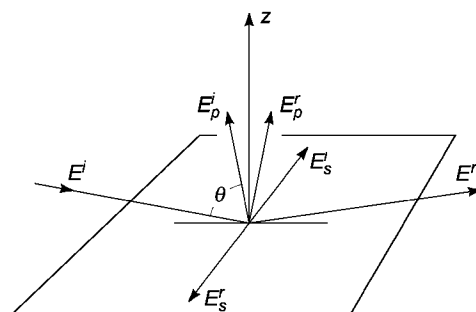


Fig. 25. Reflection geometry for irras showing the *s* and *p* components of the electric fields of the incident (*E*) and reflected (*E'*) beams (41).

The ir spectra acquired in this way are extremely sensitive to the orientation of the surface molecules. Molecules must have a significant component of a molecular vibration perpendicular to the surface to be sensed by coupling with the highly directional electric field. Molecules whose dipole moments are perfectly parallel to the surface cannot couple to the existing electric fields, and therefore, are ir transparent by this method. This selectivity of the approach for molecule dipole moments perpendicular as opposed to parallel to the surface is known as the surface selection rule of irras.

The signal-to-noise ratio of such measurements can be further improved by making the inherently single beam ftr system into a double beam system using polarization modulation. In this approach, a device which alternately passes *p*-polarized and *s*-polarized ir radiation is used to alternate the polarization of light falling on the surface. The *p*-polarized light is absorbed by not only those species that are on the surface as noted above, but also by any species in the ambient phase above the surface. The *s*-polarized light is not absorbed by the surface species; thus, it is only absorbed by species in the ambient phase above the surface. Assuming that the phase above the surface is perfectly isotropic, the difference $I_p - I_s$ should provide a spectrum due exclusively to surface species.

This approach works quite well for species at metal surfaces. It has been used extensively in recent years to ascertain information about organic thin films on metal surfaces. Of particular interest in many of these studies, and indeed the real forte of this technique, has been the determination of molecular orientation on surfaces from such studies. Few other techniques are quite so useful for unambiguously ascertaining molecular orientation.

Summary

An introduction to several of the more common methods of surface and interface analysis has been presented in this article. This treatment is certainly not comprehensive. An ever-expanding number of methods for the interrogation of surfaces and interfaces are available to the analyst. The ones chosen for discussion here were meant to be representative of methods that can answer the more general questions posed at the beginning of this article. The reader is encouraged to pursue further reading on other techniques for specific applications in the many excellent monographs on the subject of surface and interface analysis.

BIBLIOGRAPHY

1. B. W. Rossiter and R. C. Baetzold, eds., *Investigations of Surfaces and Interfaces*, Part A. *Physical Methods of Chemistry*, Vol. 9A, 2nd ed., John Wiley & Sons, Inc., New York, 1993.
2. C. R. Brundle and A. D. Baker, eds., *Electron Spectroscopy: Theory, Techniques and Applications*, Vol. 4, Academic Press, New York, 1981.
3. D. J. Connor, B. A. Sexton, and R. St. C. Smart, eds., *Surface Analysis Methods in Materials Science*, *Springer Series in Surface Sciences*, Vol. 23, Springer-Verlag, Berlin, 1992.
4. J. C. Riviere, *Surface Analytical Techniques. Monographs on the Physics and Chemistry of Materials*, Oxford University Press, New York, 1990.
5. C. L. Putzig, M. A. Leugers, M. L. McKelvy, G. E. Mitchell, R. A. Nyquist, R. R. Papenfuss, and L. Yurga, *Anal. Chem.* **66**, 26R (1994); C. L. Putzig, M. A. Leugers, M. L. McKelvy, G. E. Mitchell, R. A. Nyquist, R. R. Papenfuss, and L. Yurga, *Anal. Chem.* **62**, 270R (1990); D. R. Louder and B. A. Parkinson, *Anal. Chem.* **66**, 84R (1994); S. R. Snyder and H. S. White, *Anal. Chem.* **64**, 116R (1992); N. H. Turner and J. A. Schreifels, *Anal. Chem.* **68**, 309R (1996); G. E. McGuire, M. L. Swanson, N. R. Parikh, S. Simko, P. S. Weiss, R. J. Nemanich, D. R. Chopra, and A. R. Chourasia, *Anal. Chem.* **67**, 199R (1995).
6. J. F. O'Hanlon, *A User's Guide to Vacuum Technology*, 2nd ed., John Wiley & Sons, Inc., New York, 1989.
7. J. I. Goldstein, D. E. Newbury, P. Echlin, D. C. Joy, C. Fiori, and E. Lifshin, *Scanning Electron Microscopy and X-Ray Microanalysis*, Plenum Press, New York, 1981.
8. P. J. Goodhew and F. J. Humphreys, *Electron Microscopy and Analysis*, 2nd ed., Taylor and Francis, London, 1988.
9. L. E. Murr, *Electron and Ion Microscopy and Microanalysis: Principles and Applications*, 2nd ed., *Optical Engineering*, Vol. 29, Dekker, New York, 1991.
10. L. Reimer, *Scanning Electron Microscopy*, *Springer Series in Optical Sciences*, Vol. 45, Springer-Verlag, Berlin, 1985.
11. L. Reimer, *Transmission Electron Microscopy*, *Springer Series in Optical Sciences*, Vol. 36, 2nd ed., Springer-Verlag, Berlin, 1989.
12. J. A. Venables, D. J. Smith, and J. M. Cowley, *Surf. Sci.* **181**, 235 (1987).
13. J. A. Venables, *Ultramicroscopy* **7**, 81 (1981).
14. C. J. Chen, *Introduction to Scanning Tunneling Microscopy*, *Oxford Series in Optical and Imaging Sciences*, Vol. 4, Oxford University Press, New York, 1993.
15. R. Wiesendanger and H. J. Guntherodt, *Scanning Tunneling Microscopy III: Theory of STM and Related Scanning Probe Methods*, *Springer Series in Surface Sciences*, Vol. 29, Springer-Verlag, Berlin, 1993.
16. G. Binnig and H. Rohrer, *Rev. Mod. Phys.* **59**, 615 (1987).
17. G. A. Somorjai, *Introduction to Surface Chemistry and Catalysis*, John Wiley & Sons, Inc., New York, 1994.
18. R. Howland and L. Benatar, *A Practical Guide to Scanning Probe Microscopy*, Park Scientific Instruments, 1996.
19. D. Briggs and M. P. Seah, *Practical Surface Analysis*, 2nd ed., John Wiley & Sons, Inc., New York, 1990.

20. K. Siegbahn, *ESCA: Atomic, Molecular, and Solid State Structure by Means of Electron Spectroscopy*, Almqvist and Wiksells, Uppsala, 1967.

21. C. D. Wagner, W. M. Riggs, L. E. Davis, and J. F. Moulder, in G. E. Muilenberg, ed., *Handbook of X-Ray Photoelectron Spectroscopy*, Perkin-Elmer Corporation, Eden Prairie, Minn., 1979.

22. J. H. Scofield, *J. Electron Spectros.* **8**, 129 (1976).

23. M. Gryzinski, *Phys. Rev. A* **138**, 305 (1965).

24. A. E. Hughes and C. C. Phillips, *Surf. Interface Anal.* **4**, 220 (1982).

25. M. P. Seah and W. A. Dench, *Surf. Interface Anal.* **1**, 2 (1979).

26. G. Beamson and D. Briggs, *High Resolution XPS of Organic Polymers*, John Wiley & Sons, Inc., New York, 1992.

27. K. W. Nebesny, B. L. Maschoff, and N. R. Armstrong, *Anal. Chem.* **61**, 452A (1989).

28. L. E. Davis, N. C. MacDonald, P. W. Palmberg, G. E. Riach, and R. E. Weber, *Handbook of Auger Electron Spectroscopy*, 2nd ed., Perkin-Elmer Corporation, Eden-Prairie, Minn., 1976.

29. W. Duane and F. L. Hunt, *Phys. Rev.* **6**, 166 (1915).

30. D. E. Newbury, C. E. Fiori, R. B. Marinenko, R. L. Myklebust, C. R. Swyt, and D. S. Bright, *Anal. Chem.* **62**, 1159A, 1245A (1990).

31. R. Tertian and F. Claisse, *Principles of Quantitative X-Ray Fluorescence Analysis*, Heyden, London, 1982.

32. R. Castaing, *Ph.D. Dissertation*, University of Paris, 1951.

33. H. A. Kramers, *Philos. Mag.* **46**, 836 (1923).

34. K. F. J. Heinrich, *Electron Beam Microanalysis*, Van Nostrand, New York, 1981.

35. A. T. Bell, in J. T. Yates and T. E. Madey, eds., *Vibrational Spectroscopy of Molecules on Surfaces*, Plenum Press, New York, 1987.

36. P. R. Griffiths and J. A. de Haseth, *Fourier Transform Infrared Spectroscopy*, John Wiley & Sons, Inc., New York, 1986.

37. J. R. Ferraro and L. J. Basile, eds., *Fourier Transform Infrared Spectroscopy*, Vols. 1–3, Academic Press, New York, 1978–1982.

38. P. Kubelka and F. Munk, *Z. Tech. Phys.* **12**, 593 (1931).

39. P. Kubelka, *J. Opt. Soc. Am.* **38**, 448 (1948).

40. N. J. Harrick, *Internal Reflection Spectroscopy*, John Wiley & Sons, Inc., New York, 1967.

41. B. E. Hayden in Ref. 35.

42. R. G. Greenler, *J. Chem. Phys.* **44**, 310 (1966).

43. R. G. Greenler, *J. Chem. Phys.* **50**, 1963 (1969).

Jeanne E. emberton
University of Arizona

THERMOGRAPHY

Thermography or thermal imaging is the technique of detecting spatial and time variations of radiated heat energy and transforming them to a visible display. The equipment, called the sensor, consists of an (ir) camera connected to a television and video tape recorder. The heart of the ir camera (1) is an array of infrared detectors sensitive from visible to at least 5 μm and preferably to 14 μm wavelength (see PHOTODETECTORS). The detector array and a set of optics provide for two-dimensional spatial resolution of the thermal scene. The high speed response of the sensor results in "real time" television displays. Since thermal radiation is a function of emissivity and temperature a calibrated thermographic sensor can remotely measure the surface temperature of objects. When the camera is coupled to a spectrometer or narrow band filters the sensor is capable of spectral discrimination. Thus thermography can become both spatial radiometry and spectral imaging. The major trades-off in sensor design are typically among spatial resolution, spectral resolution, field of view, frame time and scene-temperature sensitivity.

The development of thermography began in the 1950s using either uncooled but low sensitivity bolometers or InSb photodiodes cooled to 77 K. These sensors had single detector elements, two axes scanning, large scanning mirrors, approximately 0.2 K sensitivity and a frame time exceeding 2 minutes. Adaptation to medical diagnosis was difficult because the patient was not allowed to move during the long frame time. During the 1960s linear arrays of mercury-doped germanium detectors (240 elements cooled to 35 K) were used for higher sensitivity and frame times of less than 0.1 seconds. With smaller optics, one axis scanning and a television-like display these machines were the first thermal imaging cameras. The military versions were called "forward looking infrared" or flir because heretofore military infrared with a single detector was primarily aerial reconnaissance with the airplane's motion providing one dimension of the scan. The linear array eliminated the need to move the entire airborne sensor to get a picture. During the 1970s linear detector arrays made from HgCdTe provided good sensitivity with only 77 K cooling and led to the military "Common Module" concept for lower cost, smaller size and television-like display.

From the early 1980s to present, infrared sensitive two dimensional arrays were mated to integrated circuits for signal processing and sensitivity to better than 0.03 K (see PHOTODETECTORS). These focal plane arrays of some 500 by 500 elements eliminate the need for scanning and provide good spatial resolution. Some versions have no special cooling requirements. The development trend is to increase the number of pixels to improve resolution, increase the field of view and keep the size and cost of the optics within acceptable bounds.

With reduced sensor cost the range of applications now includes thermal vision (2,3) industrial processing, industrial security, police work (3), maritime safety, airline safety and vision enhancement for night driving and flying and weather satellites. For these applications, the thermal sensor typically uses a broad spectral band to achieve highest sensitivity.

A schematic of a thermographic sensor for thermal imaging is illustrated in Figure 1. The infrared lens is often coated silicon, germanium, or an infrared glass. The ir window may also be an interference filter made of materials similar to the lens. The detector array is an integral part of the silicon integrated circuit read-out chip and contains circuits which buffer (impedance match), and switch to sequentially output the detector signals. Signal processing on the focal plane is limited to buffering, switching (multiplexing), and amplification. Typically, the focal plane signal is an analog of the photon intensity variations. The analog signals are converted to a digital format (A/D conversion) for signal processing which may consist of gain and offset corrections (for uniformity), contrast enhancement, resolution vs sensitivity tradeoff, frame addition, freeze frame (snap-shot or frame-grabbing), dithering for reduction of aliased signals, data compression for transmission, false color renditioning, and many other processing techniques. The digital signal level is typically near one volt and as such is not readily corrupted by electromagnetic interference. Although it is desirable to put the A/D converter on the focal plane for maximum noise immunity the power needed to convert at high data rates would overheat the focal plane.

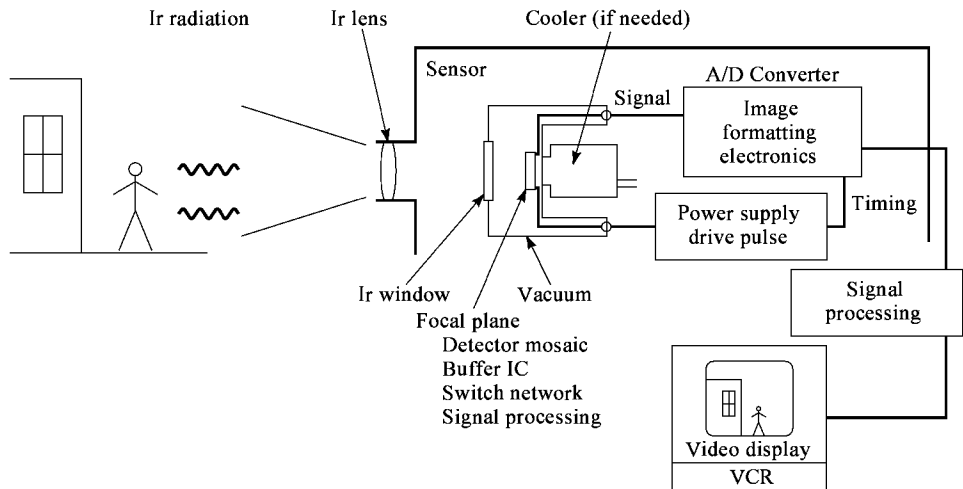


Fig. 1. Schematic for thermographic imaging. The ambient thermal radiation is imaged on the focal plane which converts the infrared to an electrical signal for display on a video monitor. Sensors with uncooled focal planes are now the size of a minicam and cost 5 to 10 times as much.

In situations where sufficient thermal radiation is available the spectral band can be narrowed (by filters or spectrometer) with only a proportional reduction in sensitivity. When the spectral band is narrowed to a particular absorption band of a given chemical substance the spatial distribution and temporal variation of that substance may be measured as was done in the visible by astronomers (scanning slit spectrographs) for many years. Infrared focal plane array technology is making possible spectrographic imaging in the infrared portion of the electromagnetic spectrum which is rich in molecular absorption lines. This will have a major impact on industrial processing where control of the chemistry is essential for cost reduction and safety. The potential applications are detection and source location of toxic gases, monitoring chemical manufacturing processes on a real-time basis, remote temperature profile sensing, automobile exhaust gas monitoring and measuring the location and degree of chemical leaks in production or storage. Spectral thermography will help detect and control toxic atmospheric gases such as CO, N_2O , SO_2 , NH_3 and ozone. A thermographic sensor with an integral gas sampling tube is a spot sensor and that with a remote heat source or retro-mirrors performs remote sensing.

Principles of Thermography

A brief review of the figures of merit (1) for thermal imaging (4) and gas detection is given to show the various trades-off required to image the thermal environment and detect atmospheric contamination.

Noise Equivalent Power. The total system electronic noise V_T may be combined with the responsivity, R_V to give the noise equivalent power equation:

$$NEP = \frac{V_T}{R_V}$$

(1)

The NEP may be written in terms of the detector element active area, A_p , the number of detector pixels elements connected for additive output n_p , the electronic noise bandwidth B and the detector element detectivity, D^* . Typically $n_p = 1$, but may be increased for improved sensitivity with an attendant loss in resolution.

$$NEP = \frac{1}{D^*} \sqrt{\frac{B A_p}{n_p}}$$

(2)

The typical D^* for a cooled infrared detector is ca 5×10^{12} cm Hz^{1/2} W and is ca 1×10^9 cm Hz^{1/2} W for an uncooled detector.

Thermographic Sensitivity. The noise equivalent temperature difference (sensitivity to scene temperature variations in degrees C) may be expressed in terms of the NEP:

$$NE\Delta T = \frac{\pi}{\Omega A_p n_p \epsilon_0} \left[\int_{\lambda_1}^{\lambda_2} \frac{dJ_s}{dT} d\lambda \right]^{-1} (NEP)$$

(3)

- where ϵ_0 = the optical efficiency of the sensor and is typically 85%
- A_p = the detector element (pixel) active area, typically 1.5×10^{-5} cm²
- Ω = the collection solid angle of the optical system ($= \pi \sin^2(\tan^{-1} 2f_n)$), typically 0.15 str
- f_n = optical fnumber, typically 2.2 or less. For fast optics $f_n = 1.0$
- J_s = spectral power density in watts/cm²cmK emitted by an element of the scene.

The integral of the temperature gradient of the spectral power density from wavelength λ_1 to λ_2 , is readily calculable using the Planck radiation law (5). Constant emissivity is assumed for equation 3.

Thermographic sensitivity for representative sensors may be calculated with equation 3. The results are summarized in Table 1 for both cooled and uncooled focal planes and for scanned and staring type sensors. Scanning sensors use a larger area detector to achieve higher sensitivity and virtual overlap of the detector geometry to suppress aliasing. The optical speed is slower ($f_n = 2$ vs $f_n = 1$) to lower the cost of optics. The D^* values are typical and refer to the cooled platinum silicide, InSb, AlGaAs superlattice or HgCdTe focal planes and the uncooled pyroelectric barium-strontium-titanate or semiconductor bolometer focal planes (1,2,6,7).

Table 1. Thermographic Sensitivity For Scanned and Staring Sensors^a

Figure of merit → Sensor type ↓	Cooled focal plane				Uncooled focal plane			
	D^* cm Hz ^{1/2} W	NEP W	$NE\Delta T$ mK	$NE\Delta T$ mK	D^* cm Hz ^{1/2} W	NEP W	$NE\Delta T$ mK	$NE\Delta T$ mK
scanned linear array $A = 2.2 \times 10^{-5}$ cm ² $B = 40$ KHz $f_n = 2.0$	5×10^{12}	19	750	80	2×10^9	470	4700	500
staring area array $A = 1.5 \times 10^{-5}$ cm ² $B = 100$ Hz $f_n = 1.0$	1×10^{10}	100	250	25	1×10^9	40	560	60

^a $\epsilon_0 = 0.85$.

In all cases, the imaging is conducted in real-time with a near standard television read-out display. This results in a 40,000 Hz bandwidth for scanning and a 100 Hz bandwidth for staring. It is possible to do frame addition and effectively reduce the bandwidth to achieve more sensitivity. The display can be repeated at the normal 60 Hz rate to avoid the appearance of flicker. Satellite infrared weather imaging thermography has a low bandwidth, resulting in a distinctive stepping motion to cloud patterns as displayed during a TV weather report. Advanced staring sensors will greatly improve weather imagery within the next decade yielding better sensitivity, resolution, and continuous movement of cloud patterns.

Thermographic imagery is considered useful when the minimum detectable temperature ($NE\Delta T$) is less than 200 mK (8). Thus the uncooled focal plane is not used in the scanning mode. The cooled area focal plane can achieve less than 5 mK but this sensitivity can not be taken advantage of because of sensor instability and limited dynamic range. Stability can be improved by periodically imaging a surface of uniform infrared irradiance, as a "reference" and electronically computing for each pixel a factor for gain correction and a difference value for offset correction. For the scanned sensor these corrections are programmed to occur at the end of each scan. For the staring sensor a low rate chopper blade is a useful, if awkward, device to achieve gain and offset correction. When the drift is slow enough, an operator briefly caps the optics or closes a shutter every hour or so to perform the correction function. The gain and offset corrections can be computed and implemented in less than 20 ms. Without correction, the random drift in the output of each detector element degrades the displayed image.

The thermographic sensor is used as a remote sensing radiometer when a reference target is imaged. It is usually necessary to correct for emissivity and atmospheric transmission to determine surface temperature with a reasonable degree of accuracy.

The dynamic range (1) is limited by the focal plane electronics which is an array of integrated circuits. The small pixels force the designer of the IC chip to use aggressive design rules to leave enough room for reasonable charge storage. Even so, the small capacitors result in a dynamic range of less than 60 dB. If for example the sensitivity is only 5 mK the dynamic range becomes 5 K and objects greater than 5 K different in temperature appear black or white on the display. The dynamic range problem is further complicated by variations in emissivity of the radiating surfaces. Since manufactured materials have a wide range in emissivity, boundaries of structured objects are usually clearly defined in thermal imagery. The user selects between higher gain (displayed sensitivity) and higher contrast (displayed dynamic range). However when the emissivity is uniform (eg, bodies of water, rural scenes) high sensitivity is desired and good stability is critical.

The resolution quality of thermal imaging is the product of the MTF or modulation transfer functions (1) for each component in the sensor system. The ideal MTF is that produced by the detector to detector spacing. Optical aberrations, photon scattering at the focal plane, shared signal charge in the detector material, electromagnetic interference and display blur combine to degrade the sensor MTF. A good sensor will have an MTF of 80% of the

optimum value at a spatial frequency given by the reciprocal of twice the detector spacing (Nyquist frequency). The detector array MTF is given by the sinc function of the spatial frequency. For a detector spacing or pitch of 50 μm the optimum MTF at a spatial frequency of 10 line pairs mm is 64%. An achievable MTF of 50% at Nyquist frequency is quite good. The minimum resolvable temperature given by equation 3 can be achieved only at low spatial frequency. Temperature resolution degrades as the image size of objects approaches detector size. Thus an imaging figure of merit, minimum resolvable temperature (MRT) is defined as the resolvable temperature difference on a display observed by a person when the bar pattern frequency matches the detector pitch, ie, at the Nyquist frequency. The MRT is typically 50% more than the NEΔT.

Thermographic Spectroscopy

The broad-band spectral capability of most thermographic sensors is generating a lot of interest for developing a low cost imaging spectrometer. The focal plane itself is sensitive from visible well into the infrared. The cooled HgCdTe array responds to 12 μm wavelength and the uncooled pyroelectric and bolometer arrays can efficiently detect from the uv to 50 μm (see PHOTODETECTORS). The spectral transmission of the window material of the detector package must be specified and the imaging optics must be designed for the specific application. Spectral discrimination is achieved by plane-wave interference, diffraction gratings, spectral wedge filters or an array of filters. Spectral dispersion is aligned parallel to a linear array so that each element is exposed to a different part of the spectrum. The linear array takes the place of the film or plate. The area arrays add the other spatial dimension and can be used for enhanced sensitivity by signal addition. When the entire array detects the same portion of the spectrum the display shows the spatial features in that spectral band. In certain cases the entire array output can be summed to achieve high sensitivity and fast response.

Although thermal imaging can be achieved by observing the irradiation of the ambient environment, spectral imaging is a different matter. Bodies, including liquids and gases emit radiation according to their temperature and emissivity and therefore act as emitting gray bodies. In addition to gray body emission substances have specific spectral emissions depending on excitation energy somehow transferred to the substance. Excitations are typically generated by electric spark, uv radiation, or localized heating. Detection in these cases is by spectral emissions. When there is no external excitation spectral discrimination by a remote spectrometer is difficult relying on second-order effects. In general, detection of small amounts of a specific material such as a toxic gas will depend on a powerful excitation source or absorption of excess ir radiation. If only internal radiation (by the material's own temperature) is available the material will be spectrally gray and will not be detected. However, an external ir radiation source as shown in Figure 2 such as a heated body can provide enough excess radiation for spectral absorption by the material and therefore spectral discrimination and identification.

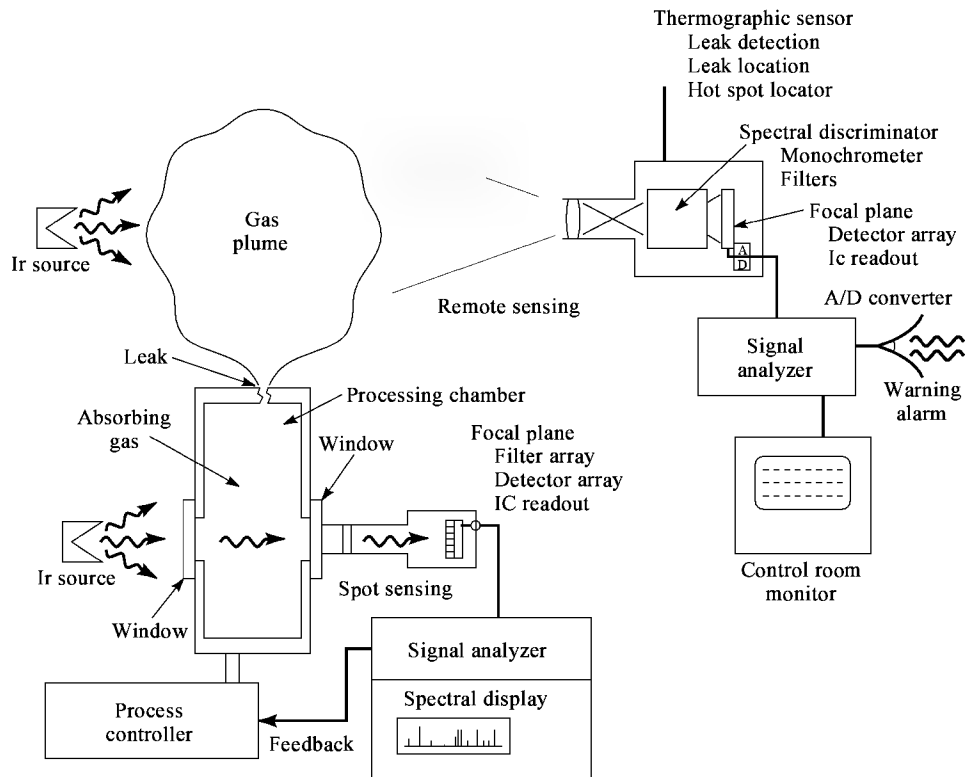


Fig. 2. Thermography combined with precise spectral discrimination provides for spatial spectroscopy for remote spectral sensing as well as "spot sensing" for process control and safety.

Spectral discrimination (9) and specific gas detection can be modeled if one assumes the gas absorbs photons of a specific wavelength exponentially with distance into the gas (Beer's law). When the absorption distance is x (cm), the incident ir power density at the detector in the spectral band pass is J ($W \text{ cm}^2$) and the power density incident on the gas is J_o , the gas concentration, C (ppm) is given by:

$$C = \frac{G_o}{x} \ln \left(\frac{J_o}{J} \right) \tag{4}$$

G_o (ppm-cm) is the normalized absorption length for a specific gas at a specific wavelength centered in specific bandwidth. Both J_o and J are referred to the collection optics. It is a measure of atomic or molecular absorption. The energy absorbed by the gas molecules is dissipated by molecular collisions and photon radiation. Even if the molecules emit the exact same wavelength, very little of the emission is in the direction of the thermographic sensor and selective absorption will be observed. The G_o value for each gas can be calculated using high resolution spectral absorption data bases by integrating the absorption of the spectral lines in the preselected bandwidth. The data for a specific gas is searched for a convenient absorption (not corrupted by other gases that may be present) band and G_o is calculated for a bandwidth consistent with spectrometer capability. For example Figure 3 shows attenuation of infrared for carbon monoxide for a 0.2-μm bandwidth centered on 4.65 μm. Since the slope is nearly constant in the range where the absorption is less than 1%, G_o is given by the slope of the curve. It is clear that the thermographic spectrometer must have high sensitivity to detect small amounts of CO in a reasonable path length of less than 1 meter.

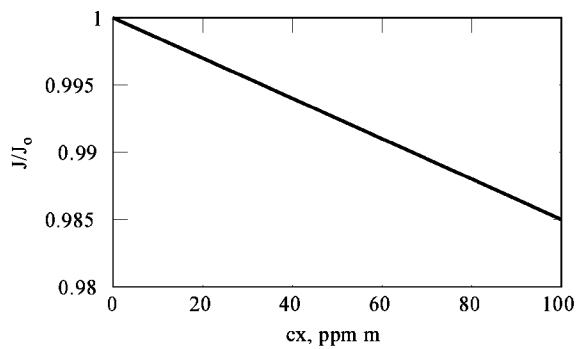


Fig. 3. Attenuation of infrared as a function of normalized concentration caused by CO absorption in the 4.6 μm band. $G_o = 6.6\text{E}3 \text{ ppm m}$, spectral bandwidth = 0.2 μm.

The detection of a specific gas (10) is accomplished by comparing the signal of the detector that is constrained to the preselected spectral band pass with a reference detector having all conditions the same except that its preselected spectral band is not affected by the presence of the gas to be detected. Possible interference by other gases must be taken into account. It may be necessary to have multiple channels or spectral discrimination over an extended spectral region to make identification highly probable. Except for covert surveillance most detection scenarios are highly controlled and identification is not too difficult.

The sensitivity equation can be developed by differentiating equation 4 with respect to J . Since the signal is proportional to J and detection is defined as to when the measured signal to noise ratio equals 3, the gas detection sensitivity (ppm) in terms of the NEP for gas detection becomes:

$$\Delta C = \frac{3G_o NEP \exp\left(\frac{x C}{G_o}\right)}{J_o A_p (n_s n_p)^{1/2} x} \tag{5}$$

This expression includes the use of detector arrays of n_p detectors with additive signals and sample addition of n_s samples to improve sensitivity. Typical sensor parameters are $J_o = 1 \text{ mW cm}^2$, $NEP = 30 \text{ pW}$, $A_p = 1.5\text{E} - 5 \text{ cm}^2$, $n_s = 60$, $n_p = 1$ for imaging and ca 600 for nonimaging gas detection.

For the case of CO detection at 4.65 μm wavelength, where $G_o = 6.6 \text{E}5 \text{ ppm-cm}$, the sensitivity as a function of x and C is shown in Figure 4. The number of signal samples n_s can be taken at a 60 Hz rate giving in this case a measuring time of 1 second. For an absorption tube length of only 50 cm the sensitivity of detection is 0.4 ppm. The specified margin of 3 in the signal to noise ratio rather than 1 improves the probability of identification from 50% to 98%.

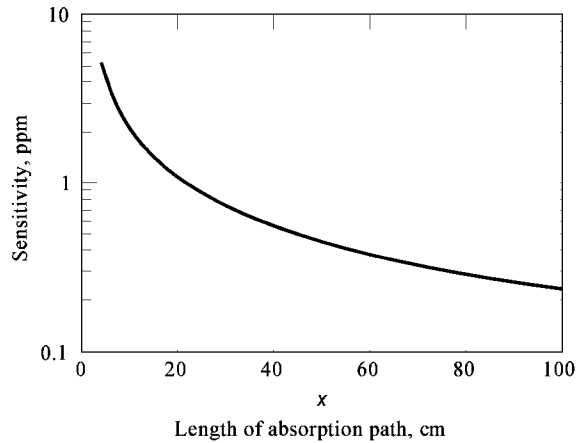


Fig. 4. Sensitivity for the detection of CO using spectral thermography as a function of absorbing path length. Detector NEP = 30 pW, $J = 1.0 \text{ mW cm}^2$, $A = 1.5\text{E} - 5 \text{ cm}^2$, samples = 60, number of detectors = 600, $G_o = 6.6\text{E}5 \text{ ppm-cm}$; center wavelength = 4.65 μm; spectral bandwidth = 0.2 μm.

An imaging thermographic sensor may be modified for spectral imaging by using a monochromator or putting a narrow band filter over the sensor aperture (11). The filter may have a constant wavelength or tapered wavelength in the form of a wedge or wheel. A schematic of this arrangement is shown in Figure 2 for the applications of industrial process control and safety. When high spectral resolution (hyperspectral) is required an interferometer spectrometer is utilized in conjunction with the thermographic sensor. At the high dispersion necessary for accurate substance identification remote sensing is very difficult because the signal level is very low. A very narrow spectral bandwidth results in a high value of G_o for gas detection. This can be alleviated somewhat by sensing two or more portions of the spectrum where absorption is expected (multispectral) and performing signal analysis for identification. The presence of several chemicals in any typical scene complicates the entire process.

Applications

Thermal Imaging. Thermography as a night vision tool in the nonmilitary sector is being used by law enforcement (3) agencies to detect the movement of suspects in the complete dark. In fact contraband discarded by suspects in flight have been detected and recovered using their latent heat signature acquired by body contact. A multipurpose commercial thermographic sensor is shown in Figure 5. Sensors are finding use by rescue services looking for survivors floating in the water at night. Since the average adult body even at rest has the energy consumption of a 100 W light bulb and emits 30% of its radiation in the 8 to 12 μm spectral band, the lightly clothed human is a bright thermographic source. All mammals are easy to detect in relatively clear weather making it easier to count livestock at night than in daylight.



Fig. 5. The Night Sight commercial thermograph made for all-weather operation uses the uncooled pyroelectric focal plane.

Courtesy of Texas Instruments Inc.

Thermal imaging is under development as nighttime vision enhancement for drivers of trucks and automobiles. When perfected the driver will have a display projected on the windshield (heads-up display) showing the forward view as seen by the infrared camera. Distant vision will be similar to daylight viewing. Humans and animals approaching the roadway will be readily seen and there will be no glare from street light or even oncoming headlights. Thermal vision in rain and fog will be better than unaided, but infrared absorption and scattering by water droplets will still greatly reduce penetration. Automobile companies project this option to be available by 2001 and will cost 5% of the average car price. Although thermal imaging is common in military aircraft, commercial airlines have been slow to introduce it because of the high cost of the airborne version. It is being used by a few airlines to help with collision avoidance and landing in light fog.

Efforts have been made to evaluate human trauma caused by disease and injury using thermal imaging. Investigations continue by medical schools to make thermography an early screening tool for the detection of kidney disease and breast cancer (12). A cancer in its early stages grows quickly creating a small amount of excess heat which appears as a bright spot on a thermograph display. Since only one breast is likely to be affected at the early stage, the radiologist looks for thermal pattern asymmetry in comparison of one breast to the other. However complications caused by blood vessels near the skin and weak response to deep lesions, have led to unacceptably high false-positive and false-negative detection as determined by the medical community. A medical thermograph is shown in Figure 6.



Fig. 6. Thermograph of a person in the 3 to 5 μm spectral band. Spatial resolution is 1 mm and temperature sensitivity is 0.1°C.

Industrial use of thermal imaging typically is the detection of thermal anomalies (13,14) such as leaking pipes and valves, overheating boilers, transformers and power lines, and friction generated heat in bearings. A thermal image of a transformer station is shown in Figure 7. These are simple heat detection applications but the imaging process quickly locates the problem in the three-dimensional industrial environment. The ir camera is scanned like a TV camera and the output signal is digitized for evaluation by a programmed computer than compares the output of each frame with a calibrated reference frame. An alarm declares when a thermal anomaly is detected. Defects in materials (15) and circuit boards (16) can be detected by the associated discontinuities in thermal conduction.

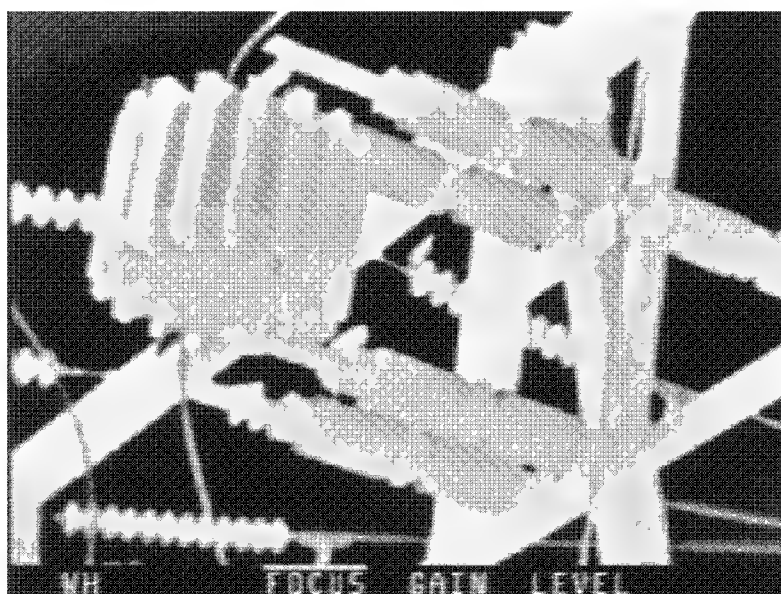


Fig. 7. Thermographic image of a power distribution station and transformer. Spatial resolution is 0.3 m-radians and temperature sensitivity is 0.08 C.

Courtesy of Texas Instruments Inc.

Chemical Gas Detection. Spectral identification of gases in industrial processing and atmospheric contamination is becoming an important tool for process control and monitoring of air quality. The present optical method uses the *ftir* (Fourier transform infrared) interference spectrometer having high resolution ($<1\text{ cm}^{-1}$) capability and excellent sensitivity (few ppb) with the use of cooled MCT (mercury–cadmium–telluride) (2) detectors. The spectral data is massive and requires normalization to remove the effects of interfering gases. The *ftir* equipment is expensive ($>100\text{K}\$$) and difficult to move around. Low cost ($<5\text{K}\$$), small size ($<10\text{ kg}$), and low power spectral thermography is under development. However, its lower spectral discrimination ($>20\text{ cm}^{-1}$) is an issue and will limit the use of the spectral thermograph to situations where the gases to be detected are known but their concentrations must be monitored. This will result in the use of well-defined spectral filtering and signal processing.

Spectral monitoring for chemical process control (17) has been demonstrated for semiconductor integrated circuit manufacture (18,19) and detection of contamination of processing gases and liquids (20). Open path or remote *ir* sensing (see Fig. 2) is used to monitor workplace gas and vapor exposures, emissions from hazardous waste sites and to track emissions along fence lines (21). Aerosol analysis by infrared spectroscopy are used to set mass detection limits and spectral quality (22). The removal of NO_x has been monitored by analysis of the by-products of the process using *ir* spectroscopy (23). *Ir* spectroscopy will be an important process control technique when the system cost is significantly reduced (24). Air quality monitoring increases in importance and U.S. Government agencies are beginning to create Guidance Documents (25,26) for the development and utilization of spectral thermographs.

BIBLIOGRAPHY

1. M. C. Dudzik, ed., *Electro-Optical Systems Design, Analysis and Testing*, Vol. 4 of J. Accetta and D. Shumaker, eds., *The Infrared and Electro-Optical Systems Handbook*, Environmental Research Institute of Michigan, 1993.
2. H. Kaplan, *Photonics Spectra*, 46 (March 1995).
3. P. W. Kruse, *Photonics Spectra*, 103 (March 1995).
4. S. B. Campana, ed., *Passive Electro-Optical Systems*, Vol. 5 of ref. 1.
5. R. B. Emmons, *Infrared Physics* **17**, 415 (1977).
6. K. H. Herrmann, *Measurement* **8**(1), 17 (Jan.–March 1990).
7. H. Kaplan, *Photonics Spectra*, 52 (June 1996).
8. B. F. Andresen, ed., *Proceedings of SPIE—The International Society for Optical Engineering* **2269**, 804 (1994).
9. F. G. Smith, ed., *Atmospheric Propagation of Radiation*, Vol. 2 of Ref. 1.
10. M. D. Schaebede and P. J. Treado, *Proc. Microscopy Society of America*, 156 (1994).
11. P. J. Treado, I. W. Levin, and E. N. Lewis, *Appl. Spectrosc.* **48**(5), 607–615 (May 1994).
12. V. A. Kalugin, A. I. Goshenko, V. S. Vetoshnikov, M. E. Belov, and P. M. Grigorishin, *Biomed. Inst. Biomed. Eng.* **23**(4), 155 (July–Aug. 1989).
13. R. J. Murry and B. F. Mitchell, *Proceedings of Annual Reliability and Maintainability Symposium (RAMS)*, 24–27 Jan. 1994, Anaheim, Calif.
14. G. J. Weil and R. J. Graf, *Second Natl. Spec. Conf. Civ. Eng. Appl. Remote Sens. Georg. Inf. Syst.*, ASCE, New York, pp. 189–198 (1991).
15. M. P. Luong, *Nucl. Eng. Des.* **158**(2–3), 363–376 (Sept. 1995).
16. S. Zannoli, *Elettronica-Oggi* (148), 98–100, 106 (Oct. 15, 1992).
17. I. D. Aggarwal, S. Farquharson, and E. Koglin, eds., *Proceedings of SPIE—The International Society for Optical Engineering* **2367**, 243p (1995).
18. C. J. Gamsky, P. M. Dentinger, G. R. Howes, and J. W. Taylor, *Proceedings of SPIE—The International Society for Optical Engineering* **2438**, 143–152 (1995).
19. K. Nishikawa, K. Ono, M. Tuda, T. Oomori, and K. Namba, *Jpn. J. Appl. Physics* **34**(7A), 3731–3736 (July 1995).
20. *Symposium on On-Line Monitoring Institute of Environmental Sciences—Proceedings*, Annual Technical Meeting 1995, Institute of Environmental Sciences, Mount Prospect, Ill., 1995, 106p.
21. M. S. Malachowski, S. P. Levine, G. Herrin, R. C. Spear, M. Yost, and Z. Yi, *J. Air Waste Management Assoc.* **44**(5), 673–682 (May 1994).
22. D. T. Allen and E. Palen, *J. Aerosol Sci.* **20**(4), 441–455 (1989).
23. A. Mizuno, K. Shimizu, A. Chakrabarti, I. Dascalescu, and S. Furuta, *IEEE Trans. Industry Applic.* **31**(5), 957–964 (Sept.–Oct. 1995).
24. O. Oehler, *Sensor Rev.* **15**(3), 14–16 (1995).
25. G. Russwurm and J. W. Childers, *FTIR Open Path Monitoring, Guidance Document*, EPA Contract 68-D5-0049, ManTech Environmental Technology Inc., Research Park, N.C., 1996.
26. B. McClenny, *Compendium Method TO-16*, U.S. EPA, National Exposure Research Laboratory, Research Triangle Park, N.C., 1996.

Sebastian Borrello
Texas Instruments

UREA

Urea [57-13-6] was discovered in urine by Rouelle in 1773 and first synthesized from ammonia (qv) and cyanic acid by Woehler in 1828. This was the first synthesis of an organic compound from an inorganic compound, and it dealt a deathblow to the vital-force theory. In 1870, urea was produced by heating ammonium carbamate in a sealed tube.

Properties

Urea can be considered the amide of carbamic acid, NH₂COOH, or the diamide of carbonic acid, CO(OH)₂. At room temperature, urea is colorless, odorless, and tasteless. Properties are shown in Tables 1–4. Dissolved in water, it hydrolyzes very slowly to ammonium carbamate (1) and eventually decomposes to ammonia and carbon dioxide (qv). This reaction is the basis for the use of urea as fertilizer (qv).

Table 1. Properties of Urea

Property	Value
melting point, °C	135
index of refraction, <i>n</i> ²⁰ _D	1.484, 1.602
density, d ²⁰ ₄ , g cm ³	1.3230
crystalline form and habit	tetragonal, needles or prisms
free energy of formation, at 25°C, J mol ^a	197.150
heat of fusion, J g ^a	251 ^b
heat of solution in water, J g ^a	243 ^b
heat of crystallization, 70° aqueous urea solution, J g ^a	460 ^b
bulk density, g cm ³	0.74
specific heat, J (kg·K) ^a	
at 0°C	1.439
50	1.661
100	1.887
150	2.109

^a To convert J to cal, divide by 4.184.

^b Endothermic.

^c Exothermic.

Table 2. Properties of Saturated Aqueous Solutions of Urea

Temperature, °C	Solubility in water, g 100 g solution	Density, g cm ³	Viscosity mPa·s(–cP)	H ₂ O vapor pressure, kPa ^a
0	41.0	1.120	2.63	0.53
20	51.6	1.147	1.96	1.73
40	62.2	1.167	1.72	5.33
60	72.2	1.184	1.72	12.00
80	80.6	1.198	1.93	21.33
100	88.3	1.210	2.35	29.33
120	95.5	1.221	2.93	18.00
130	99.2	1.226	3.25	0.93

^a To convert kPa to mm Hg, multiply by 7.5.

Table 3. Properties of Saturated Solutions in Urea in Ammonia^a

Temperature, °C	Urea in solution, wt %	Vapor pressure of solution, kPa ^b
0	36	405
20	49	709
40	68	952
60	79	1094
80	84	1348
100	90	1267
120	96	507

^a Ref. 2.

^b To convert kPa to atm, divide by 101.3.

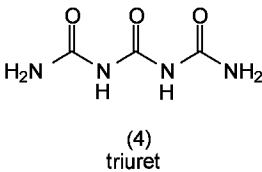
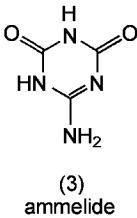
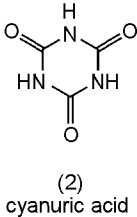
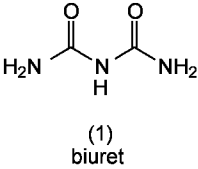
Table 4. Properties of Saturated Solutions of Urea in Methanol and Ethanol^a

Temperature, °C	Methanol		Ethanol	
	Urea, wt %	Density, g cm ³	Urea, wt %	Density, g cm ³
20	22	0.869	5.4	0.804
40	35	0.890	9.3	0.804
60	63	0.930	15.0	0.805

^a Ref. 3.

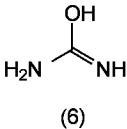
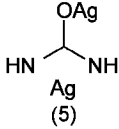
Commercially, urea is produced by the direct dehydration of ammonium carbamate, NH₂COONH₄, at elevated temperature and pressure. Ammonium carbamate is obtained by direct reaction of ammonia and carbon dioxide. The two reactions are usually carried out simultaneously in a high pressure reactor. Recently, urea has been used commercially as a cattle-feed supplement (see FEEDS AND FEED ADDITIVES). Other important applications are the manufacture of resins (see AMINO RESINS AND PLASTICS), glues, solvents, and some medicinals. Urea is classified as a nontoxic compound.

At atmospheric pressure and at its melting point, urea decomposes to ammonia, biuret (1), cyanuric acid (qv) (2), ammelide (3), and triuret (4). Biuret is the main and least desirable by-product present in commercial urea. An excessive amount (>wt%) of biuret in fertilizer-grade urea is detrimental to plant growth.

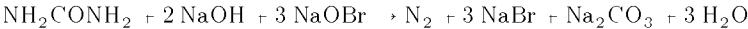


Urea acts as a monobasic substance and forms salts with acids (4). With nitric acid, it forms urea nitrate, CO(NH₂)₂·HNO₃, which decomposes explosively when heated. Solid urea is stable at room temperature and atmospheric pressure. Heated under vacuum at its melting point, it sublimes without change. At 180–190°C under vacuum, urea sublimes and is converted to ammonium cyanate, NH₄OCN (5). When solid urea is rapidly heated in a stream of gaseous ammonia at elevated temperature and at a pressure of several hundred kPa (several atm), it sublimes completely and decomposes partially to cyanic acid, HNCO, and ammonium cyanate. Solid urea dissolves in liquid ammonia and forms the unstable compound urea–ammonia, CO(NH₂)₂·NH₃, which decomposes above 45°C (2). Urea–ammonia forms salts with alkali metals, eg, NH₂CONHM or CO(NHM)₂. The conversion of urea to biuret is promoted by low pressure, high temperature, and prolonged heating. At 10–20 MPa (100–200 atm), biuret gives urea when heated with ammonia (6–7).

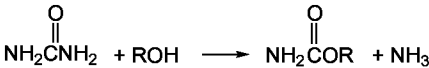
Urea reacts with silver nitrate, AgNO₃, in the presence of sodium hydroxide, NaOH, and forms a diargentate derivative (5) of a pale-yellow color. Sodium hydroxide promotes the change of urea into the imidol form (6):



which then reacts with silver nitrate. Oxidizing agents in the presence of sodium hydroxide convert urea to nitrogen and carbon dioxide. The latter reacts with sodium hydroxide to form sodium carbonate (8):

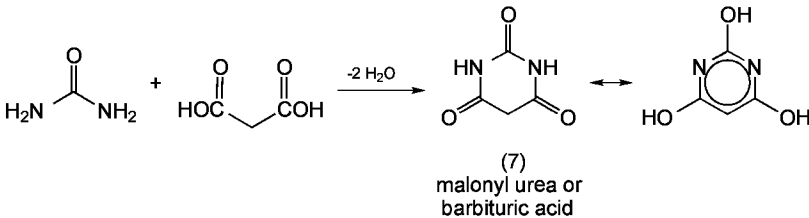


The reaction of urea with alcohols yields carbamic acid esters, commonly called urethanes (see URETHANE POLYMERS):



Urea reacts with formaldehyde compounds such as monomethylolurea, NH₂CONHCH₂OH, dimethylolurea, HOCH₂NHCONHCH₂OH, and others, depending on the mol ratio of formaldehyde, to urea and upon the pH of the solution. Hydrogen peroxide and urea give a white crystalline powder, urea peroxide, CO(NH₂)₂·H₂O₂, known under the trade name of Hypersol, an oxidizing agent.

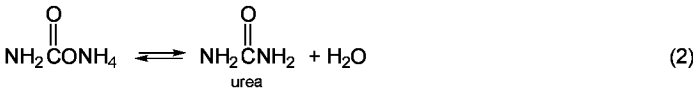
Urea and malonic acid give barbituric acid (7), a key compound in medicinal chemistry (see also HYPNOTICS, SEDATIVES, AND ANTICONVULSANTS):



Manufacture

Urea is produced from liquid NH₃ and gaseous CO₂ at high pressure and temperature; both reactants are obtained from an ammonia-synthesis plant. The latter is a by-product stream, vented from the CO₂ removal section of the ammonia-synthesis plant. The two feed components are delivered to the high pressure urea reactor, usually at a mol ratio >2.5 : 1. Depending on the feed mol ratio, more or less carbamate is converted to urea and water per pass through the reactor.

The formation of ammonium carbamate and the dehydration to urea take place simultaneously, for all practical purposes:



Reaction 1 is highly exothermic. The heat of reaction at 25°C and 101.3 kPa (1 atm) is in the range of 159 kJ/mol (38 kcal/mol) of solid carbamate (9). The excess heat must be removed from the reaction. The rate and the equilibrium of reaction 1 depend greatly upon pressure and temperature, because large volume changes take place. This reaction may only occur at a pressure that is below the pressure of ammonium carbamate at which dissociation begins or, conversely, the operating pressure of the reactor must be maintained above the vapor pressure of ammonium carbamate. Reaction 2 is endothermic by ca 31.4 kJ/mol (7.5 kcal/mol) of urea formed. It takes place mainly in the liquid phase; the rate in the solid phase is much slower with minor variations in volume.

The dissociation pressure of pure carbamate has been investigated extensively (10–12) and the average values are shown in Table 5.

Table 5. Vapor Pressure of Pure Ammonium Carbamate at which Dissociation Begins

Temperature, °C	kPa ^a
40	31
60	106
80	314
100	861
120	2,130
140	4,660
160	9,930
180	15,200; 19,300 ^b
200	20,300; 36,500 ^b

^a To convert kPa to atm, divide by 101.3.

^b The value has been extrapolated because, at temperatures above 170°C, the rate of reaction 2 rapidly increases and it is difficult to determine the carbamate vapor pressure owing to the formation of water and urea and the consequent lowering of the partial pressure of ammonium carbamate.

Ammonium Carbamate. Ammonium carbamate is a white crystalline solid which is soluble in water (2). It forms at room temperature by passing ammonia gas over dry ice. In an aqueous solution at room temperature, it is slowly converted to ammonium carbonate, (NH₄)₂CO₃, by the addition of one mol of water. Above 60°C, the ammonium carbonate solution reverts to carbamate solution, and at 100°C, only carbamate is present in the solution. Above 150°C, ammonium carbamate loses a mol of water and forms urea. The specific heat of solid ammonium carbamate is given in Table 6. Ammonium carbamate melts at ca 150°C, and has a heat of fusion of ca 16.74 kJ/mol (4.0 kcal/mol). The conversion of carbamate to urea begins at <100°C. To obtain an appreciable amount of urea at 100°C requires 20–30 h. The rate of conversion increases with increasing temperature (13–15); at 185°C, ca 50% of the ammonium carbamate is converted to urea in ca 30 min.

Table 6. Specific Heat of Solid Ammonium Carbamate

Temperature, °C	J (g·K) ⁻¹
20	1.67

60	1.92
100	2.18
140	2.43
180	2.59

^a To convert J to cal, divide by 4.184.

Conversion at Equilibrium. The maximum urea conversion at equilibrium attainable at 185°C is ca 53% at infinite heating time. The conversion at equilibrium can be increased either by raising the reactor temperature or by dehydrating ammonium carbamate in the presence of excess ammonia. Excess ammonia shifts the reaction to the right side of the overall equation:



Water, however, has the opposite effect. Actual equilibrium constants at various temperatures are given in Table 7. A detailed study of the effect of pressure on urea conversion is given in Ref. 17.

Table 7. Reaction Equilibrium Constant^a

Temperature, °C	Reaction equilibrium constant, K
140	0.695
150	0.850
160	1.075
170	1.375
180	1.800
190	2.380
200	3.180

^a Ref. 16.

Processing

At this time over 95% of all new urea plants are licensed by Snamprogetti, Stamicarbon, or Toyo Engineering. SNAM utilizes thermal stripping while STAC (Stamicarbon) and Toyo use CO₂ stripping. Only these three processes are, therefore, covered in detail. Process flow sheets are included for others at the end of this section.

As of the end of 1996, about 70 SNAM plants, 125 STAC, and 7 Toyo stripping plants have been built. Currently STAC will design plants for over 3000 MTD, SNAM about 2800 MTD, and Toyo about 2300 MTD as single train units.

STAC, with their current new design (pool reactor), feel the only size limitation will be vessel size. Unless both the vessel fabricator and the intended plant site are "on water", a 4-m-diameter is the maximum that can be transported.

The urea produced is normally either prilled or granulated. In some countries there is a market for liquid urea–ammonium nitrate solutions (32% N). In this case, a partial-recycle stripping process is the best and cheapest system. The unconverted NH₃ coming from the stripped urea solution and the reactor off-gas is neutralized with nitric acid. The ammonium nitrate solution formed and the urea solution from the stripper bottom are mixed, resulting in a 32–35 wt % solution. This system drastically reduces investment costs as evaporation, finishing (prill or granulation), and wastewater treatment are not required.

Snamprogetti Thermal Stripping Process. The Snamprogetti process is outlined in Figure 1. Initially SNAM utilized NH₃ as the stripping agent. Owing to the high solubility of NH₃ in the synthesis liquid, an overload of NH₃ occurred in the downstream recirculation sections of the plant. At this time only heat is supplied to the stripper to remove unreacted NH₃ and CO₂. Because of the high NH₃ ratio, it is still necessary to have two recirculation sections.

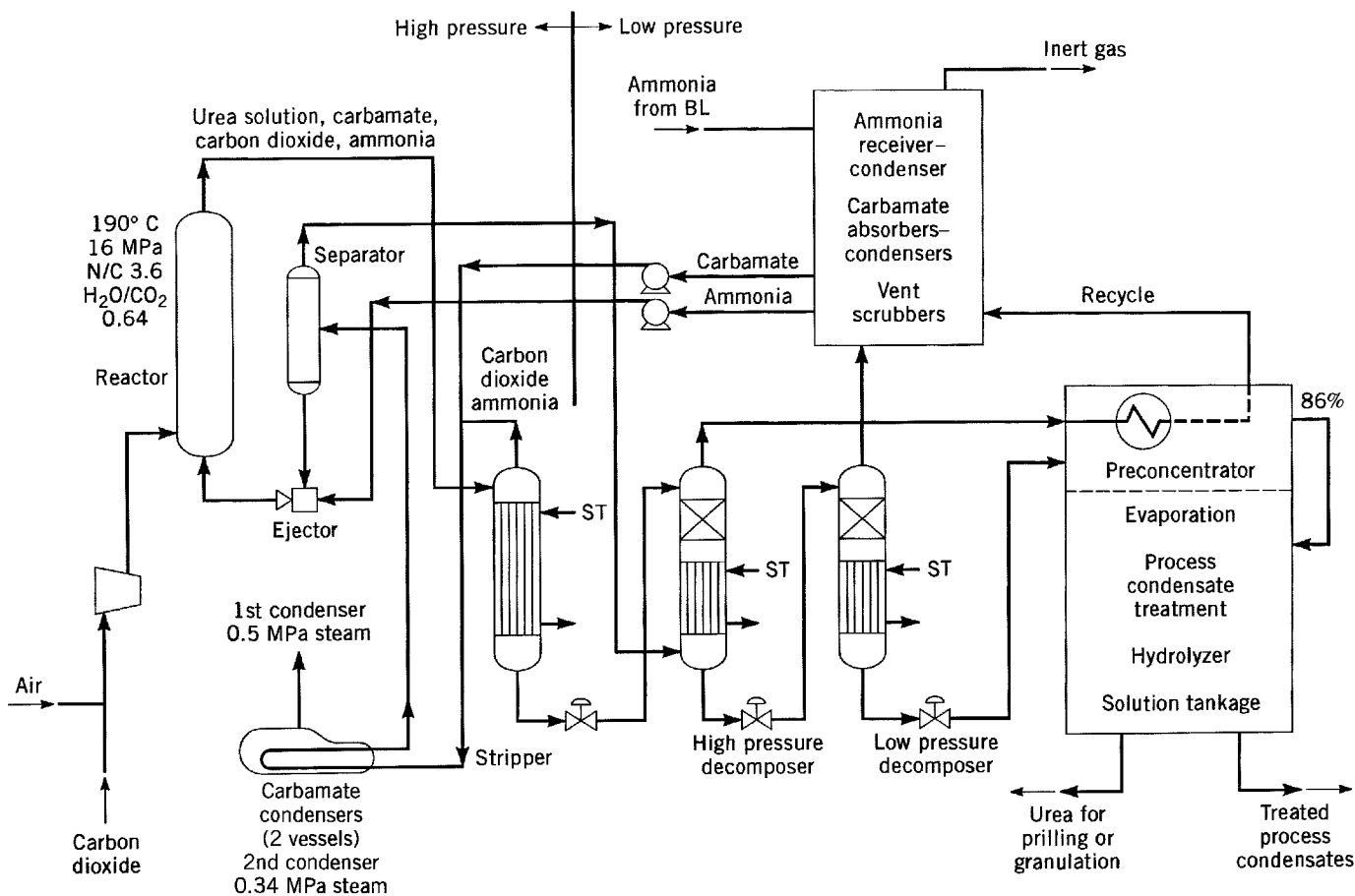


Fig. 1. Snamprogetti thermal stripping urea process. BL = battery limits.

The synthesis recycle loop has the stripped gas going to two high pressure carbamate condensers in series and to a high pressure separator and then back to the reactor. The flow is maintained by using an NH₃-driven liquid-liquid ejector. The reactor is operated at 15 MPa (150 bar) with a NH₃-CO₂ molar feed ratio of 3.5. The stripper is a falling-film type and since high temperatures (200–210°C) are required for efficient thermal stripping, stainless steel tubing is not suitable. Titanium was initially used, but it also was not satisfactory because of erosion near the bottom. At this time a bimetallic tube of zirconium and 25-22-2 stainless steel is used. The zirconium is corrosion-free and the only problem is the difficulty in getting proper welds and separation of the two layers at the bottom ends. Fabrication mistakes have been the only source of problems with this vessel to date.

The stripper off-gas going to the high pressure carbamate condensers also contains the carbamate recovered in the medium and low pressure recirculation sections. Both of these systems are similar to those shown in the total-recycle process.

The plant is designed with an excellent heat-exchange system to keep overall steam required to a minimum.

The plant wastewater containing NH₃ and urea is subjected to a desorption-hydrolysis operation to recover almost all the NH₃ and urea. In some plants, this water can then be used for boiler feedwater.

The urea solution is evaporated in a two-stage system (99.8%) if the final product is prills, and a single-stage system (+95%) if granules are to be provided.

Stamicarbon CO₂ Stripping Process. In the early 1960s, Stamicarbon introduced the first stripping process. One of the main improvements was the reduction of steam required per ton of urea to 0.8–0.95 from the 1.8 ton required in the conventional total-recycle process (Fig. 2). Steam usage and electric power required can be varied depending on the final design.

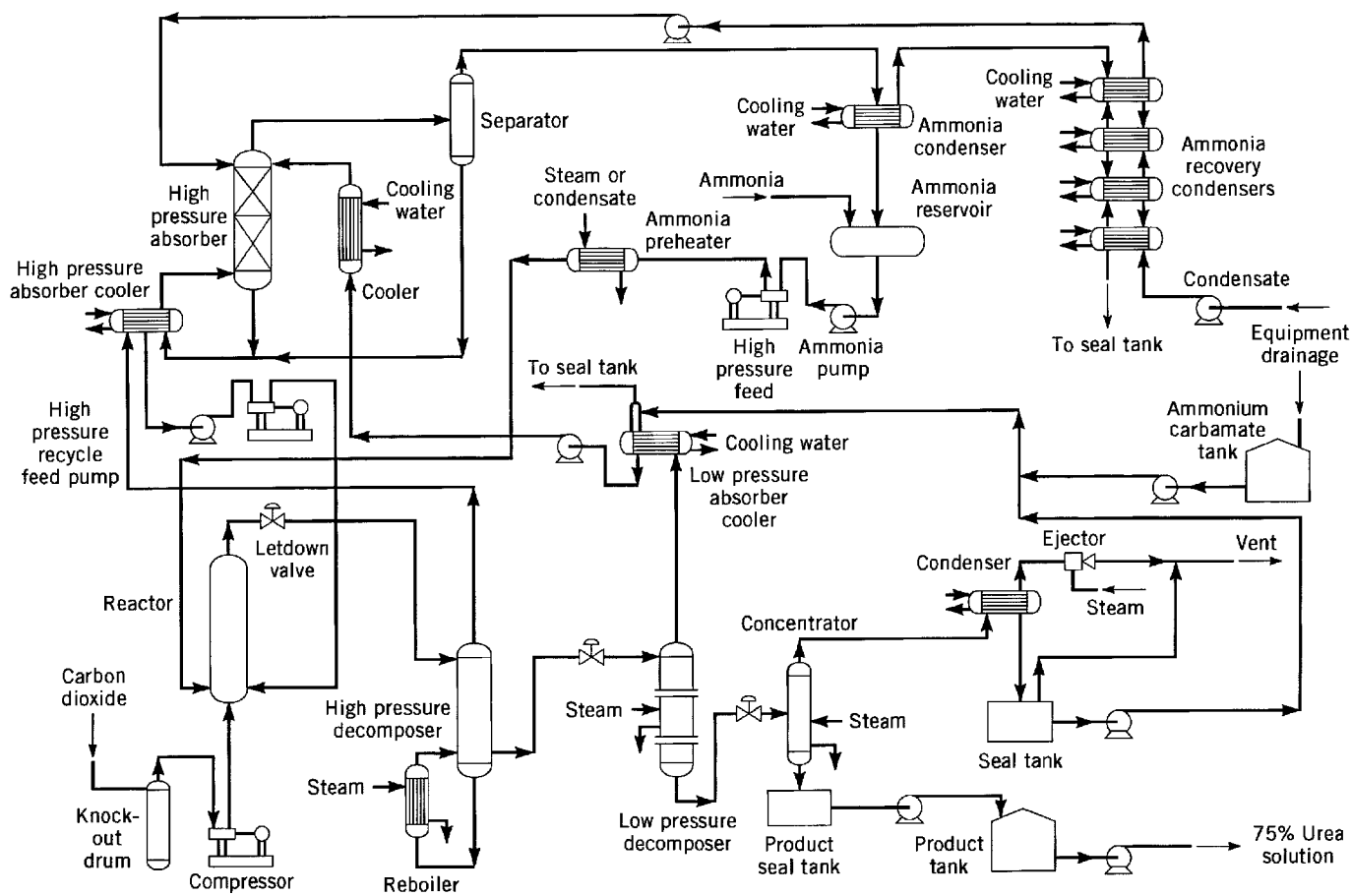


Fig. 2. Typical total-recycle urea process.

The Stamicarbon process is described in Figures 3–7. The synthesis section of the plant consists of the reactor, stripper, high pressure carbamate condenser, and a high pressure reactor off-gas scrubber. In order to obtain a maximum urea yield per pass through the reactor, a pressure of 14 MPa (140 bar) and a 2.95 : 1 $\text{NH}_3\text{--CO}_2$ molar ratio is maintained. The reactor effluent is distributed over the stripper tubes (falling-film type shell and tube exchanger) and contacted by the CO_2 countercurrently. This causes the partial NH_3 pressure to decrease and the carbamate to decompose.

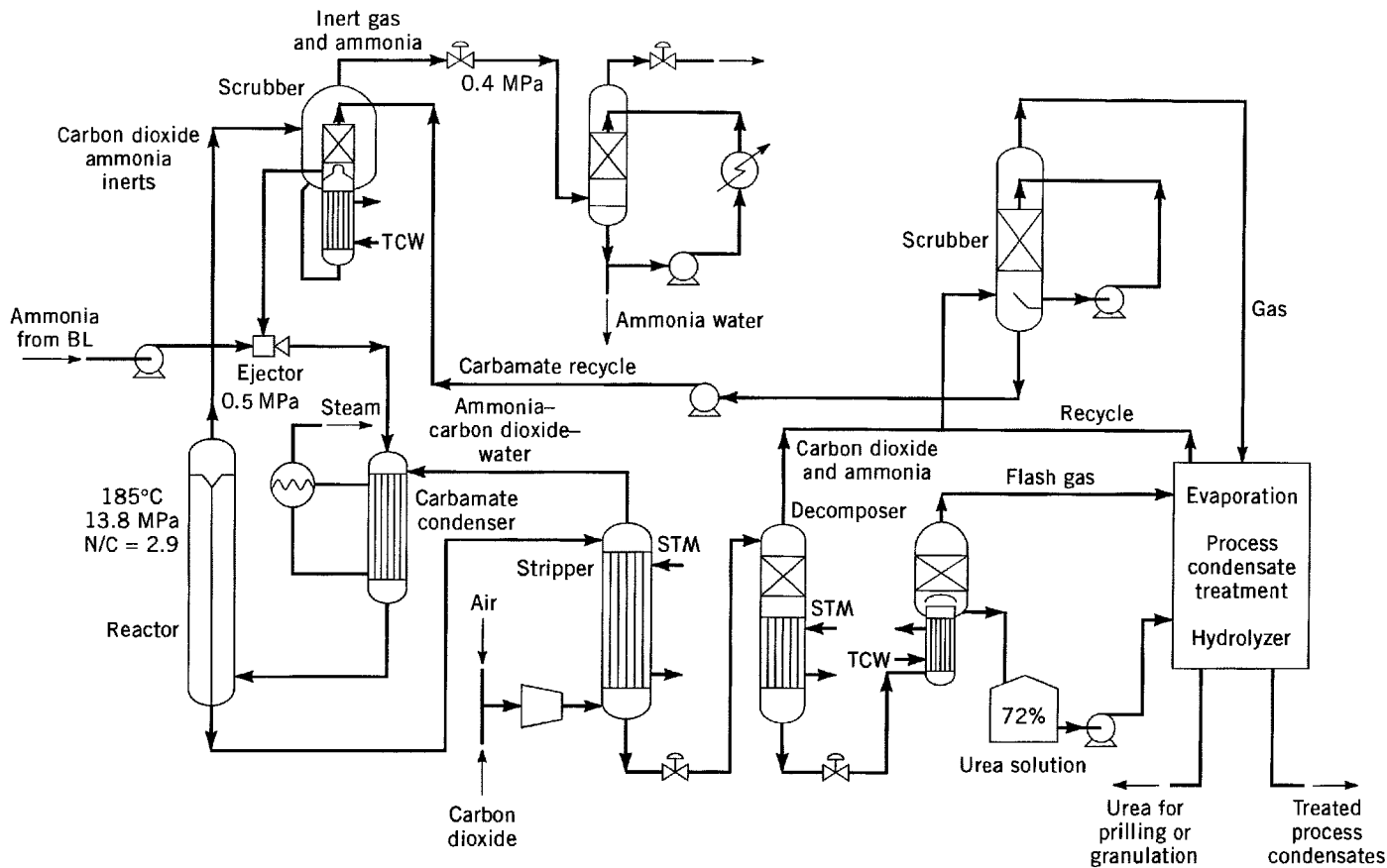


Fig. 3. Stamicarbon CO_2 stripping process. TCW = tempered cooling water.

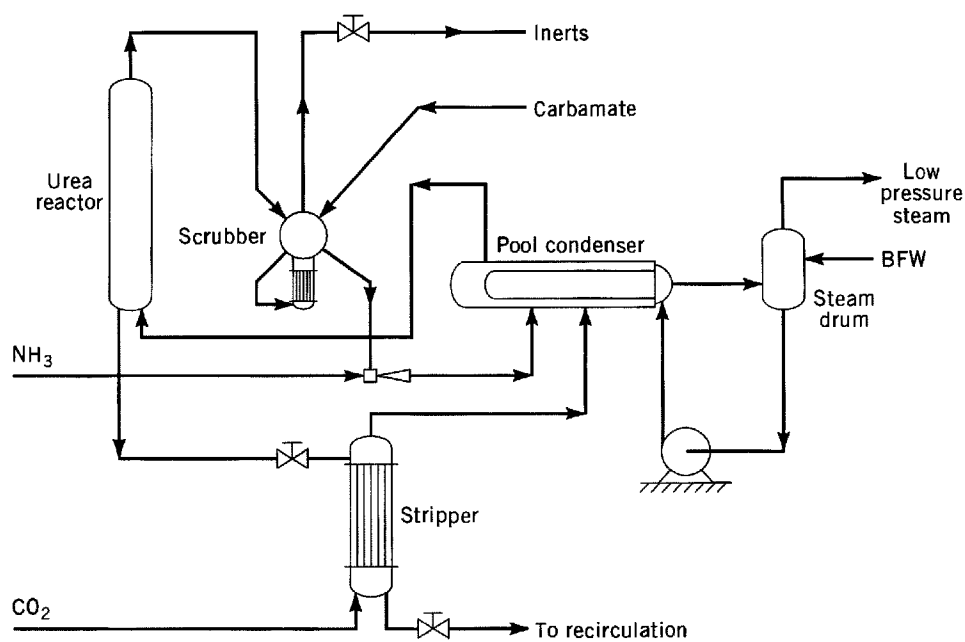


Fig. 4. Pool condenser type synthesis section for Stamicarbon CO₂ stripping process.

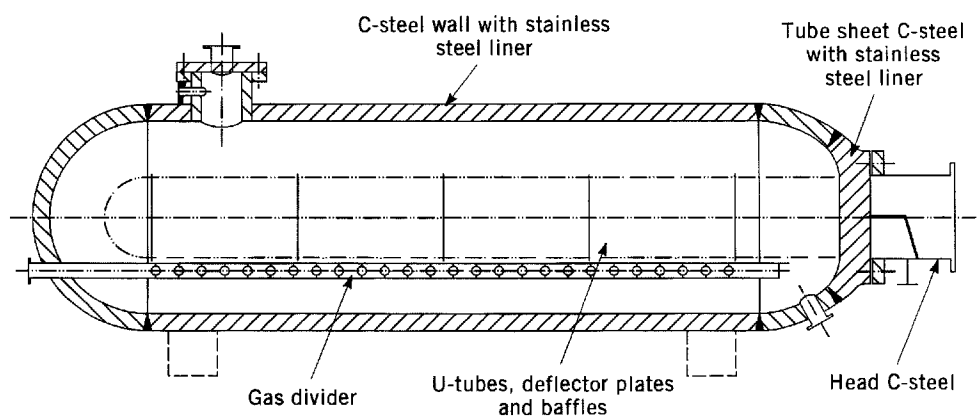


Fig. 5. Pool condenser vessel for Stamicarbon CO₂ stripping process.

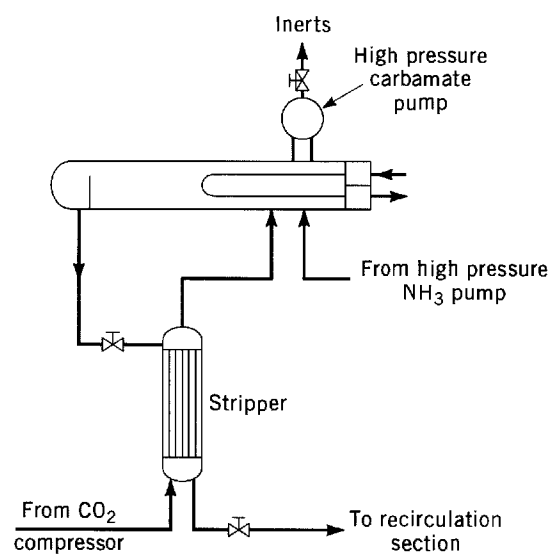


Fig. 6. Pool reactor synthesis section for Stamicarbon CO₂ stripping process.

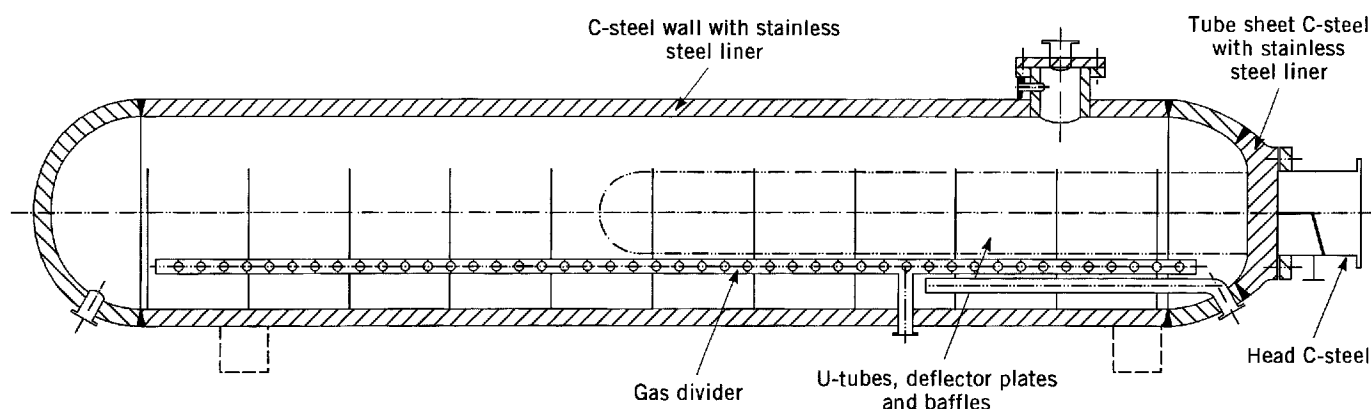


Fig. 7. Pool reactor vessel for Stamicarbon CO₂ stripping process.

The urea solution out of the stripper bottom flows to a single-stage low pressure recirculation section (0.4 MPa, 4 bar). The stripper off-gas is sent to the carbamate condenser.

In this condenser, part of the stripper off-gases are condensed (the heat of condensation is used to generate low pressure steam). The carbamate formed and noncondensed NH₃ and CO₂ are put into the reactor bottom and conversion of the carbamate into urea takes place. The reactor is sized to allow enough residence time for the reaction to approach equilibrium. The heat required for the urea reaction and for heating the solution is supplied by additional condensation of NH₃ and CO₂. The reactor which is lined with 316 L stainless steel, contains sieve trays to provide good contact between the gas and liquid phases and to prevent back-mixing. The stripper tubes are 25-22-2 stainless steel. Some strippers are still in service after almost 30 years of operation.

The noncondensable gases introduced in the CO₂ (ie, passivation air) and part of the unreacted NH₃ and CO₂ goes to the high pressure scrubber, which consists of a shell-and-tube exchanger in the bottom portion and a packed bed in the upper part. In the lower part, most of the NH₃ and CO₂ is condensed, the heat of condensation is dissipated in tempered cooling water. In the upper part, the gases leaving the bottom section are countercurrently contacted with carbamate solution returned from recirculation. The scrubber off-gas, containing nitrogen, oxygen, and very small amounts of NH₃ and CO₂ are vented to the atmosphere after passing through an absorber.

The carbamate solution from the scrubber flows to a high pressure ejector. The NH₃ feed pressure induces enough head to convey the carbamate solution from the scrubber to the carbamate condenser.

As mentioned before, because of design and operating conditions (ie, NH₃ CO₂ ratio, pressure, temperature, reactor volume), only one recirculation stage is required (0.4 MPa) (4 bar). On expansion, a large portion of the carbamate left in the urea solution from the stripper decomposes. The remaining solution passes through a rectifying column, heater, and separator. The gases formed go to a carbamate condenser and are pumped via the high pressure carbamate pump (either reciprocating or centrifugal) to the high pressure scrubber.

The urea solution is then evaporated to 99.8% for prilling (2 stages) or plus 95% for granulation (1 stage).

The Stamicarbon wastewater system consisting of two desorbers, hydrolyzer, hydrolyzer heater, reflux condenser, desorber heat exchanger, and a wastewater cooler is very efficient. Also, in many plants, as the water contains less than 1 ppm of NH₃ and of urea, it can be used as cooling water make-up, or boiler-feed water.

In 1994 Stamicarbon introduced a pool condenser in the synthesis section (see Figs. 4 and 5). This allowed a 34% decrease in reactor volume and a 45% decrease in carbamate heat-exchange area, thus reducing costs considerably for equipment, structural steel, and construction.

In late 1997, DSM (Stamicarbon's parent company) will start up a new plant utilizing, the next step, a pool reactor (see Figs. 6 and 7). In 1999 this process should be offered for licensing.

In order to be 100% safe from a hydrogen explosion (sources: passivation air, CO₂), a hydrogen removal system is installed before the CO₂ passivation air enters the stripper.

Toyo Engineering-ACES Process. The synthesis section of the ACES process (Fig. 8) consists of a reactor, a stripper, two carbamate condensers, a scrubber and operates at 17.5 MPa (175 bars). The reactor is operated at 190°C with a NH₃ CO₂ ratio of 4.0 (mol/mol). Liquid NH₃ is fed directly into the reactor by a centrifugal ammonia pump. Gaseous CO₂ is sent from the centrifugal CO₂ compressor to the bottom section of the falling-film type stripper.

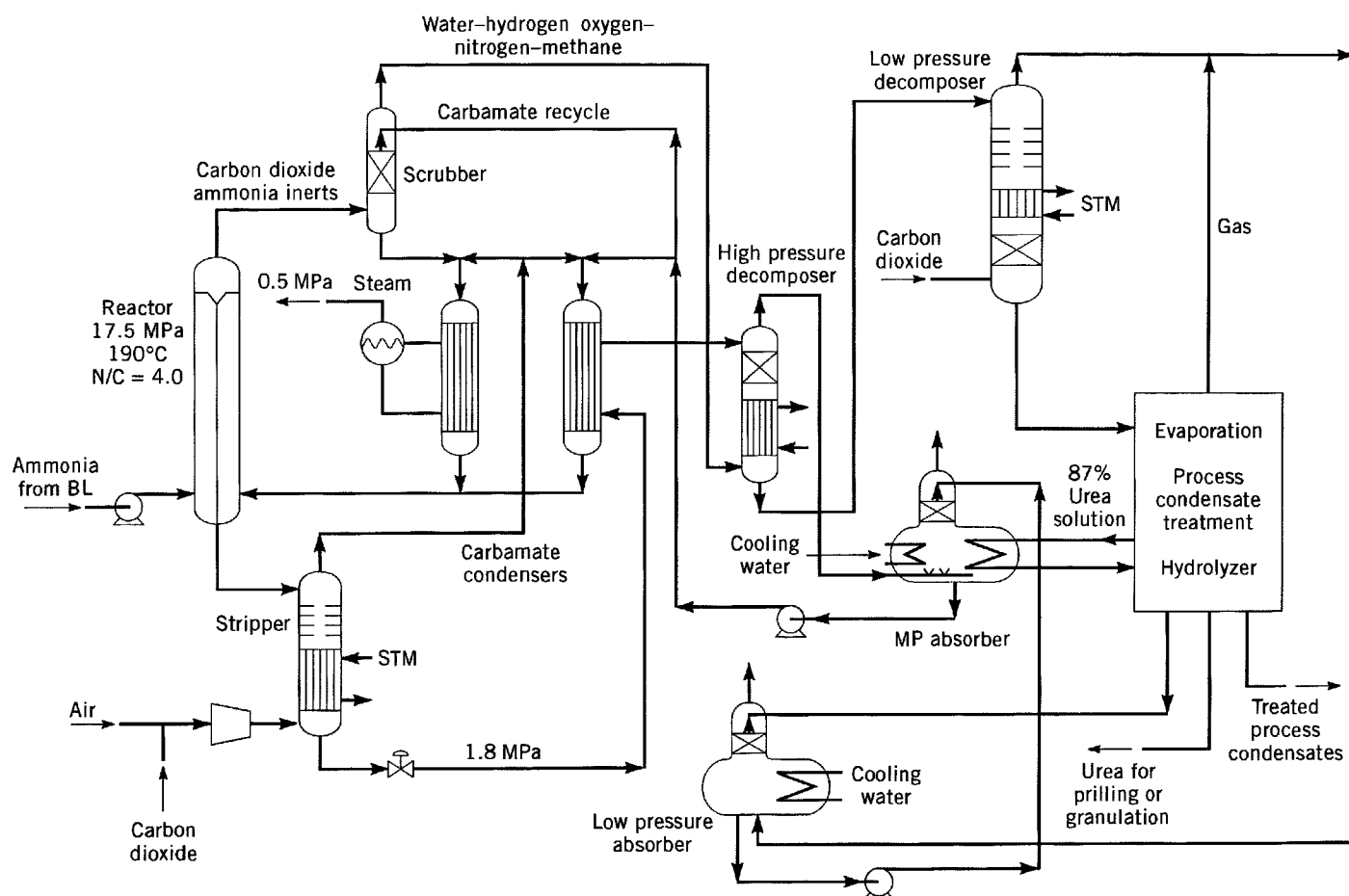


Fig. 8. TEC ACES process. MP = medium pressure; BL = battery limits.

The stream from the reactor consisting of a mixture of urea, unconverted ammonium carbamate, excess water, and NH_3 , is fed into the top of the stripper. The ACES stripper utilizes a ferrite-austenite stainless steel, as do the carbamate condensers. The reactor and scrubber are constructed with 316 L urea-grade stainless steel.

The stripper has two functions: The upper part contains trays and separates excess NH_3 in the feed (which ensures effective CO_2 stripping in the falling film type lower section). The solution then passes through the falling-film section where the ammonium carbamate is decomposed and separated by CO_2 stripping and steam heating. This overhead gas mixture is fed to the parallel carbamate condensers where the gaseous mixture is condensed and absorbed into the solutions from the scrubber and the 1.8 MPa (18 bar) absorber. One carbamate condenser is used to generate low pressure steam 0.45–0.5 MPa (4.5–5 bar) and the other to heat the urea solution from the stripper bottoms. The gas-liquid stream from the condenser is recycled back to the reactor by gravity flow. The inert stream from the top of the reactor is purged to the scrubber for recovery of the NH_3 and CO_2 .

The urea solution leaving the stripper bottom contains about 12 wt% of NH_3 and is further purified in the 1.8 MPa (18 bar) and 0.2 MPa (2 bar) recovery sections of the plant. The resultant NH_3 and CO_2 separated in the decomposers is absorbed and returned to the synthesis section by the high pressure centrifugal carbamate pump.

The urea solution stream is then fed to the vacuum concentrator unit which operates at 17.3 kPa (130 mm Hg abs) and produces 88.7 wt% urea. It then goes to either two-stage evaporators if prills are made, or a single-stage unit for granule production.

As in all the processes, the process condensate and all other sources of waste urea- NH_3 -water contamination go to a waste recovery unit which includes a urea hydrolyzer. The final water discharge is then below 3–5 ppm of NH_3 and urea.

Other Processes. Flow sheets for typical partial-recycle process and typical once-through urea process are given in Figures 9 and 10, respectively.

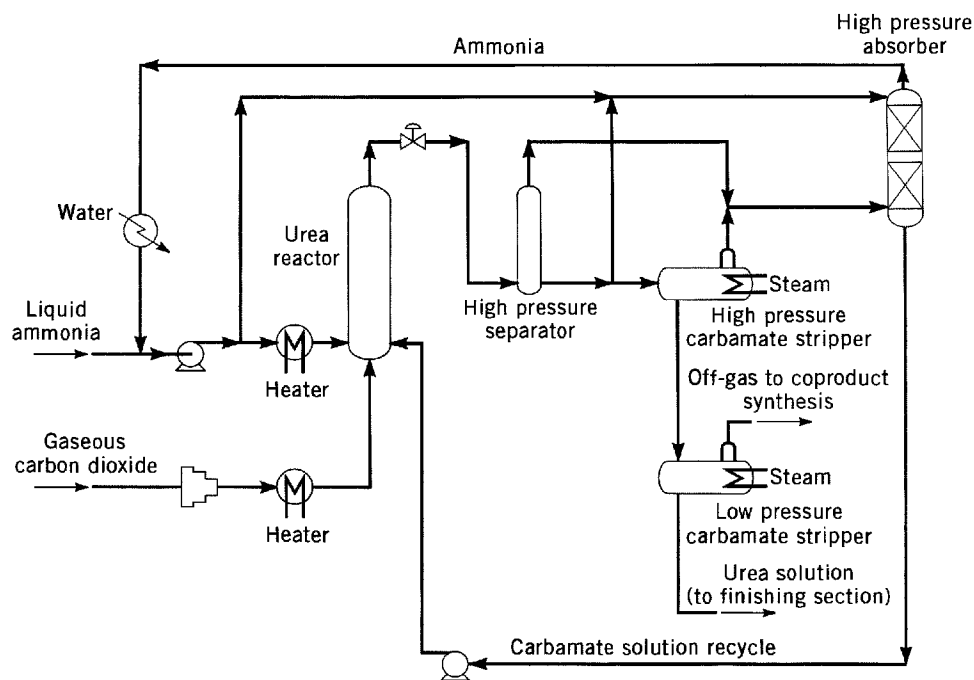


Fig. 9. Typical partial-recycle process.

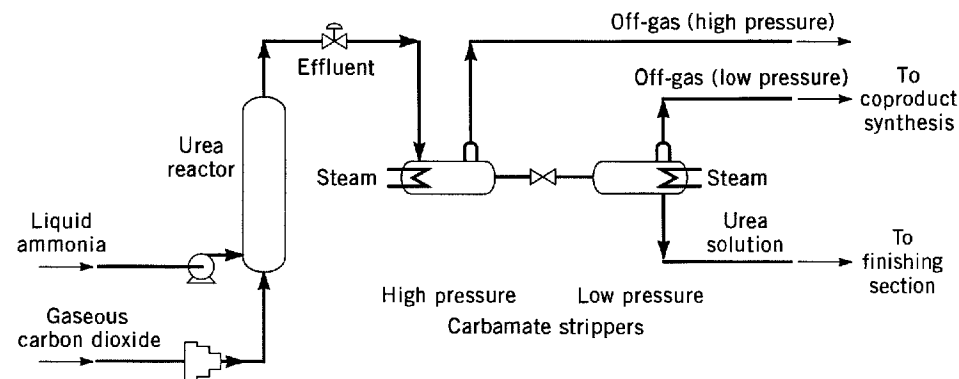


Fig. 10. Typical once-through urea process.

Finishing Processes

Urea processes provide an aqueous solution containing 70–87% urea. This solution can be used directly for nitrogen-fertilizer suspensions or solutions such as urea–ammonium nitrate solution, which has grown in popularity recently (18). Urea solution can be concentrated by evaporation or crystallization for the preparation of granular compound fertilizers and other products. Concentrated urea is solidified in essentially pure form as prills, granules, flakes, or crystals. Solid urea can be shipped, stored, distributed, and used more economically than in solution. Furthermore, in the solid form, urea is more stable and biuret formation less likely.

Prilling. The manufacture of prills is rapidly decreasing owing to both environmental problems and product quality as compared to granules. In a prilling plant the urea solution from the recovery section is evaporated in two stages to +99.8% strength. It is then pumped to the top of a 50–60m cylindrical concrete tower where it is fed into a spinning bucket containing many (+2000) small holes. The emerging small liquid droplets solidify as they fall and are cooled by a forced or induced draft air flow. The very fine dust that is formed and exits at the top of the tower with the air flow is an environmental problem. Many recovery systems have been tried (ie, water baths, dust collection, etc), but none have been very successful. The quality problem is that prill size must necessarily be small in order to obtain proper solidification and cooling in the fall height that is practical (50–60). Generally, both the crushing and impact strength of the prill is much less than for a granule. This causes many problems in handling both at the plant and in shipping. Stamicarbon has introduced a "seeding" system that has improved prill strength. Also, formaldehyde can be added as it not only improves the crushing strength, but is suppresses the caking tendency in storage. If it were not for environmental considerations, prilling would still be a cheaper option than granulation in a small-scale marketing area (ie, not on a global scale).

Granulation. Almost all new plants produce granules and the Hydro-Agri process is used in the majority of plants (Fig. 11). This process was developed by NSM of Holland many years ago. They have no plant size limitation and will design a single-train unit for production over 3000 MTD. The C&I Girdler drum system has been very successful, but cannot compete in today's market because of restrictions in the train size. Toyo has successfully developed a spout-fluid technology (Figs. 12–14.) Three plants are in operation and others are in either the design or construction stage. Stamicarbon also will license a fluid-bed large-scale single-train plant that is somewhat similar to the Hydro-Agri design (Fig. 15).

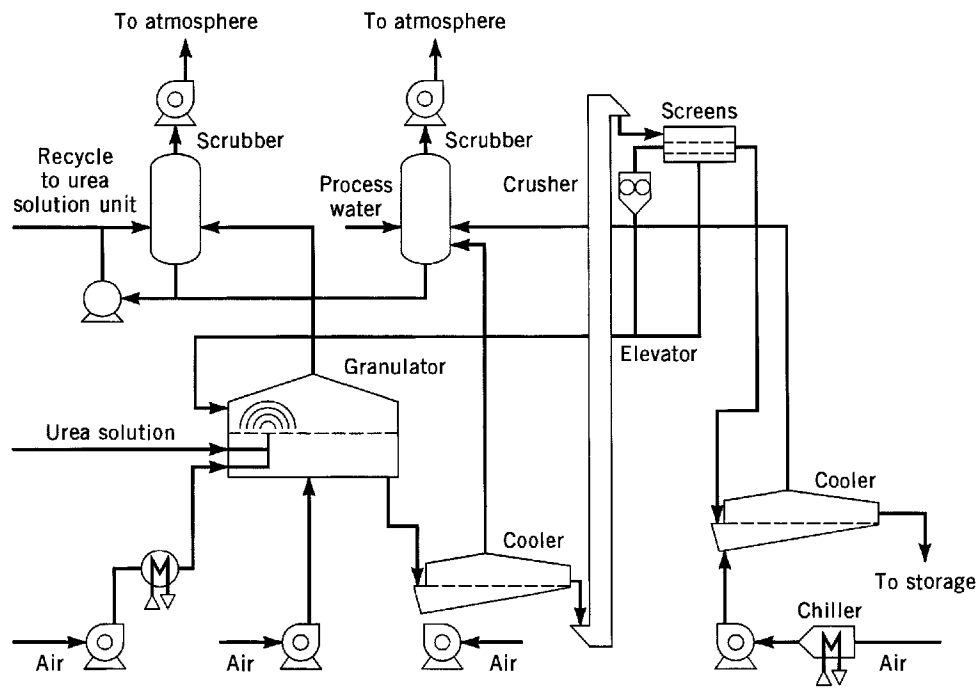


Fig. 11. Hydro-Agri urea granulation process.

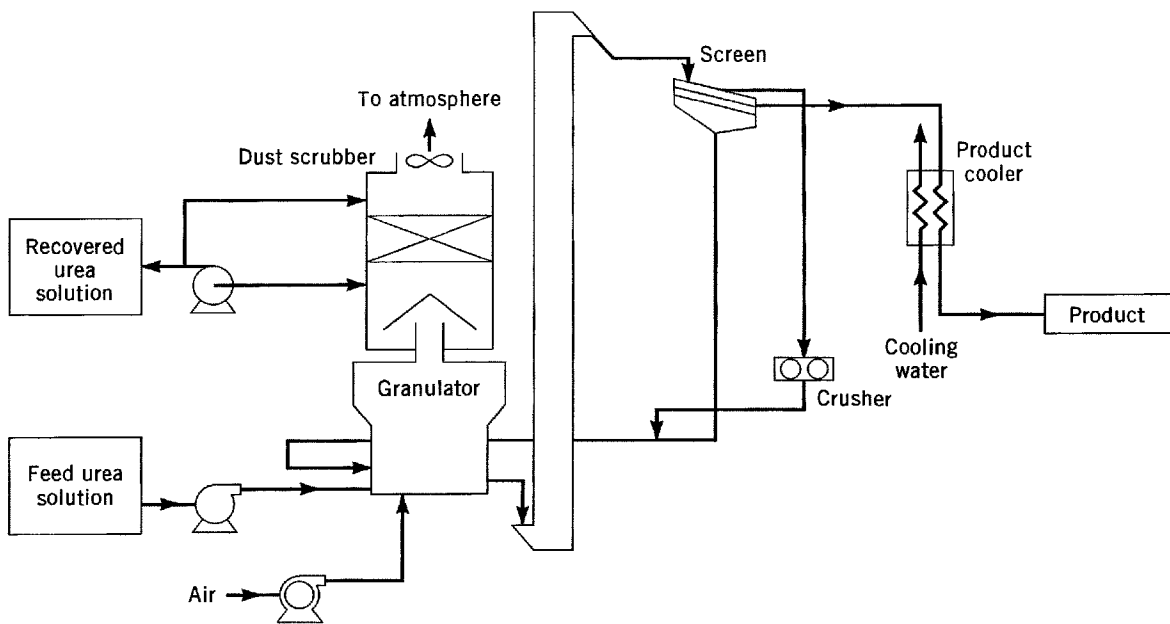


Fig. 12. TEC urea granulation process.

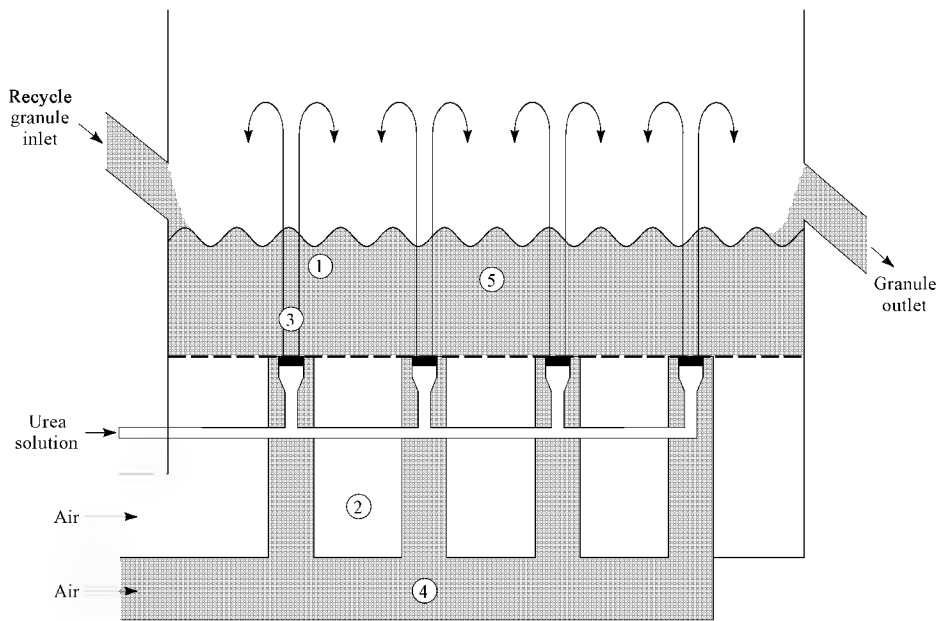


Fig. 13. Multistage spout-fluid-bed reactor. 1, spouted bed; 2, perforated plate; 3, spray nozzle; 4, air header; 5, fluidized bed.

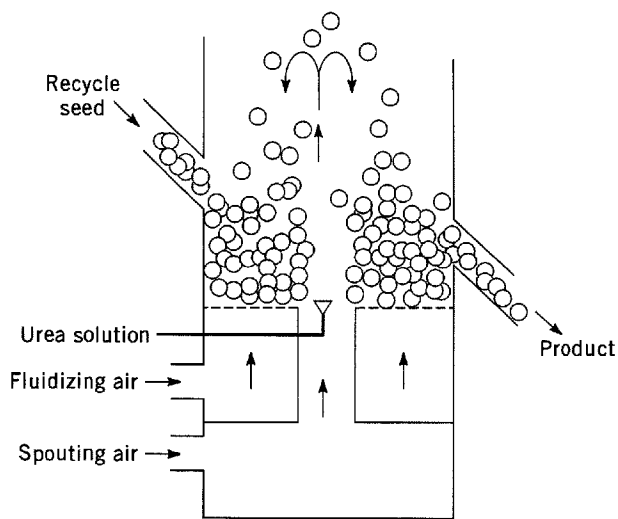


Fig. 14. Single spout-fluid-bed granulator.

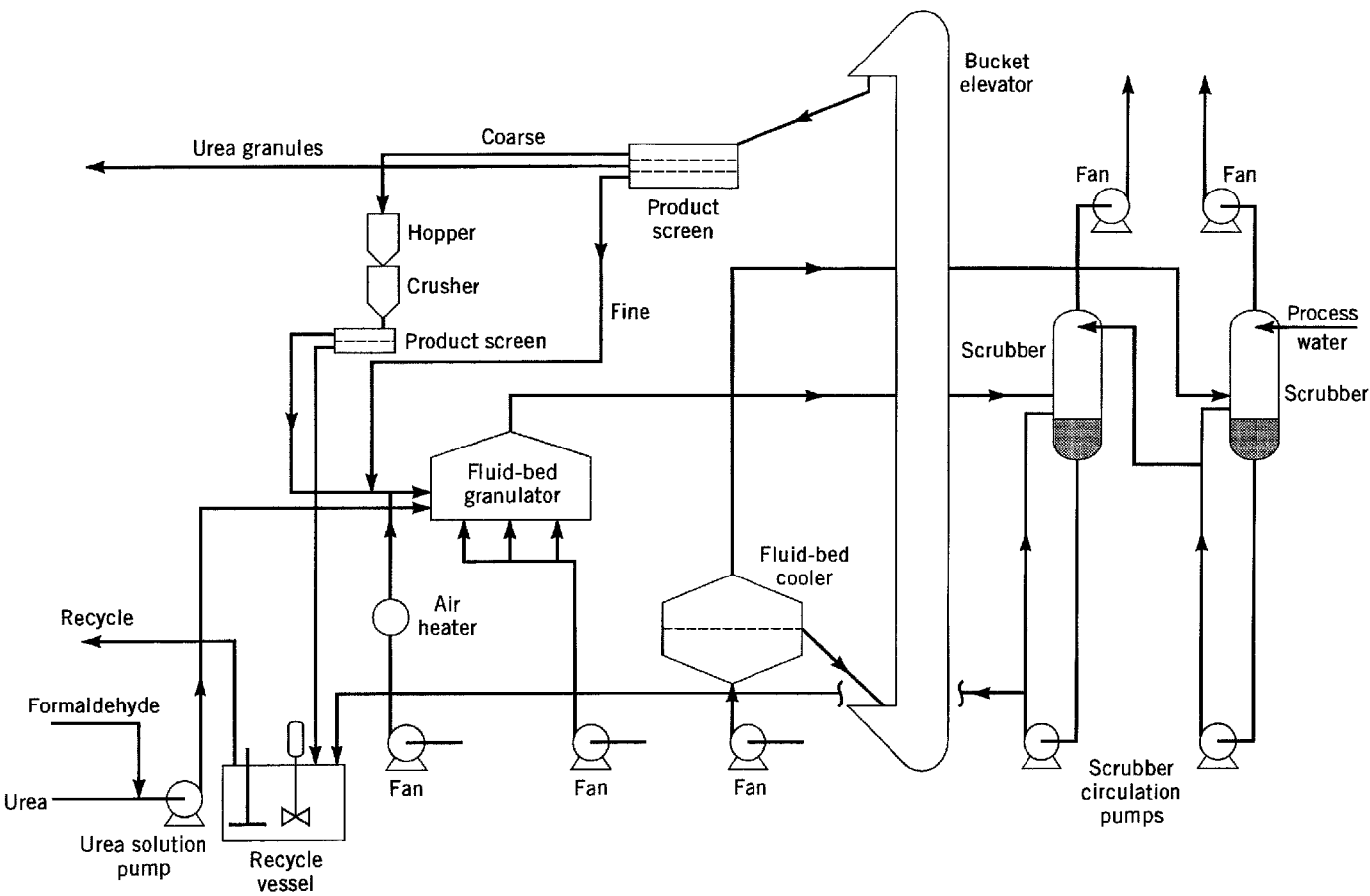


Fig. 15. Stamicarbon urea granulation.

Wastewater Treatment

Under the pressure of progressively more stringent government regulations with regard to permissible levels of residual NH_3 and urea content in wastewaters, the fertilizer industry made an effort to improve wastewater treatment (see also WATER, SEWAGE).

For each mol of urea produced in a total-recycle urea process, one mol of water is formed. It is usually discharged from the urea concentration and evaporation section of the plant. For example, a 1200 t d plant discharges a minimum of 360 t d of wastewater. With a barometric condenser in the vacuum section of the evaporation unit, the amount of wastewater is even higher. Small amounts of urea are usually found in wastewaters because of entrainment carry-over.

The problem in reducing the NH_3 and urea content in the wastewaters to below 100 ppm is because it is difficult to remove one in the presence of the other. The wastewater can be treated with caustic soda to volatilize NH_3 . However, in a more efficient method, the urea is hydrolyzed to ammonium carbamate, which is decomposed to NH_3 and CO_2 ; the gases are then stripped from the wastewater.

All process licensors also feature wastewater treatment systems. Stamicarbon guarantees the lowest NH_3 -urea content and has plants in operation confirming the low NH_3 -urea (1 ppm NH_3 -1 ppm urea). This water is very satisfactory to use as boiler feed water. See Figures 16 and 17 for this system.

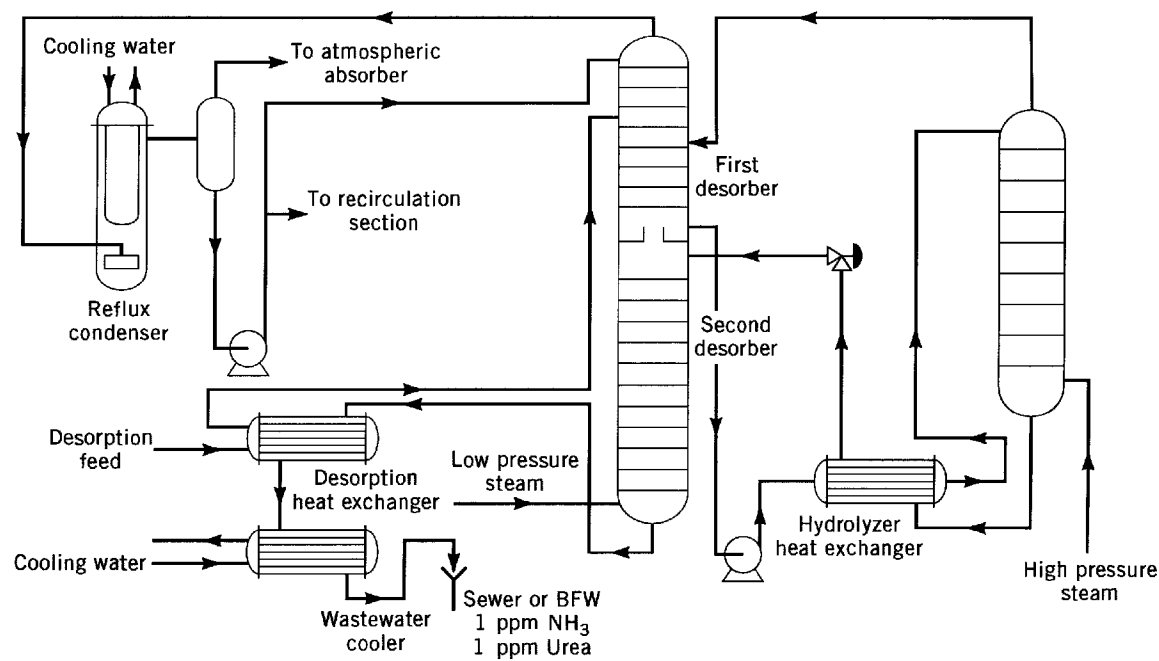


Fig. 16. Simplified flow sheet of wastewater treatment.

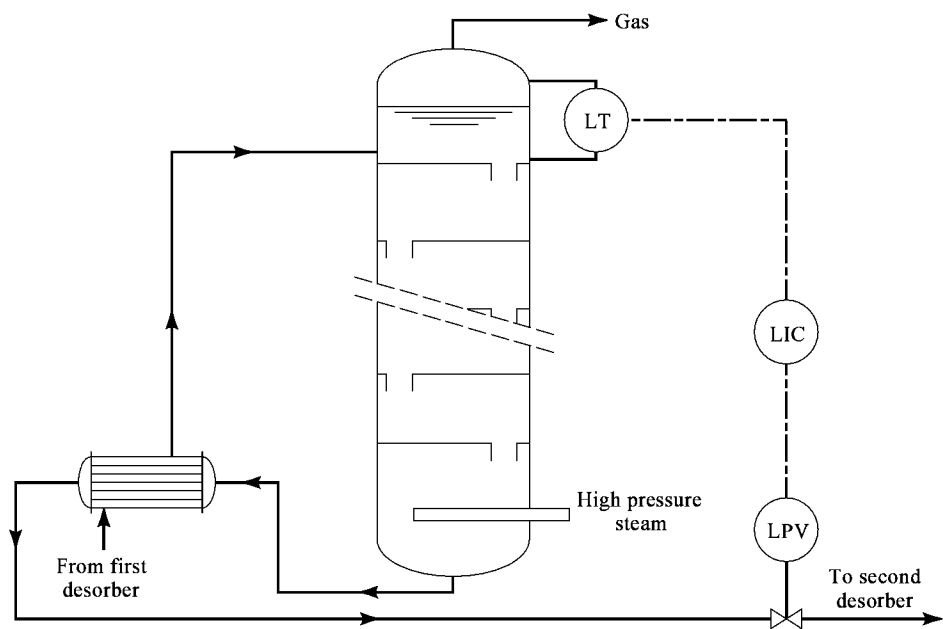


Fig. 17. Countercurrent hydrolyzer.

Economic Aspects

The U.S. urea production from 1976 to 1980 (18,19) is given in Table 8. During this period, urea became the most important solid nitrogen fertilizer. Urea solution production almost doubled from 1977 to 1978, proving its importance as a liquid fertilizer.

Table 8. U.S. Urea Production, 10⁶ t^a

Year	Fertilizer		Other	Total
	Solid	Solution		
1970	1.090	1.278	0.580	2.948
1974	1.302	1.221	0.916	3.439
1976 ^b	2.207	1.094	0.779	4.080
1977	2.435	1.337	0.829	4.601
1978	2.868	2.376	0.446	5.690
1979	3.888	1.964	0.498	6.350
1980 ^b	3.921	2.583	0.599	7.103
1996 ^c	5.707	1.982	—	7.689

^a Refs. 18, 19.
^b Annual growth rate >11.5%
^c U.S. capacity.

In 1980, urea consumption in other areas had declined, but not enough to detract from the overall bright future for the material, which had a total

U.S. production of 7.103×10^6 metric tons in 1980 in plants with a nameplate capacity of 7.1 t. Estimated urea production in the United States for 1981 is lower than the 1980 production by ca 3^o %, because of a decline in exports from 1.8 t in 1980 to ca 1.45 t in 1981 (20). Nameplate capacity of urea plants in the United States and Canada is projected to be ca 9.075×10^6 t yr by 1984 (21). World capacity should be ca 95.150×10^6 t yr (21). World capacity increased from 27.200×10^6 t yr in 1970 to ca 77.300×10^6 t yr in 1980 (22).

Uses

Solid urea containing 0.8–2.0 wt^o % biuret is primarily used for direct application to the soil as a nitrogen-release fertilizer. Weak aqueous solutions of low biuret urea (0.3 wt^o % biuret max) are used as plant food applied to foliage spray.

Mixed with additives, urea is used in solid fertilizers of various formulations, eg, urea–ammonium phosphate (UAP), urea–ammonium sulfate (UAS), and urea–phosphate (urea + phosphoric acid). Concentrated solutions of urea and ammonium nitrate (UAN) solutions (80–85 wt^o %) have a high nitrogen content but low crystallization point, suitable for easy transportation, pipeline distribution, and direct spray application.

Urea is also used as feed supplement for ruminants, where it assists in the utilization of protein. Urea is one of the raw materials for urea–formaldehyde resins. Urea (with ammonia) pyrolyzes at high temperature and pressure to form melamine plastics (see also CYANAMIDES). Urea is used in the preparation of lysine, an amino acid widely used in poultry feed (see AMINO ACIDS; FEEDS AND FEED ADDITIVES, PET FOODS). It also is used in some pesticides.

Partially polymerized resins of urea are used by the textile industry to impart permanent-press properties to fabrics (see also TEXTILES, FINISHING).

The consumption of urea for urea–formaldehyde resins has decreased in recent years because of the new findings about the toxicity of formaldehyde slowly released by the resin.

Reagent-grade urea is used in some pharmaceutical preparations. In these applications, urea must meet the purity specifications issued by the ACS.

Clathrates

Urea has the remarkable property of forming crystalline complexes or adducts with straight-chain organic compounds. These crystalline complexes consist of a hollow channel, formed by the crystallized urea molecules, in which the hydrocarbon is completely occluded. Such compounds are known as clathrates. The type of hydrocarbon occluded, on the basis of its chain length, is determined by the temperature at which the clathrate is formed. This property of urea clathrates is widely used in the petroleum-refining industry for the production of jet aviation fuels (see AVIATION AND OTHER GAS TURBINE FUELS) and for dewaxing of lubricant oils (see also PETROLEUM, REFINERY PROCESSES). The clathrates are broken down by simply dissolving urea in water or in alcohol.

BIBLIOGRAPHY

"Urea" in *ECT* 1st ed., Vol. 14, pp. 458–466, by F. A. Wolff and D. J. O'Flynn, E. I. du Pont de Nemours & Co., Inc.; "Urea" in *ECT* 2nd ed., Vol. 21, pp. 37–56, by I. Mavrovic, Consulting Engineer; "Urea" in *ECT* 3rd ed., Vol. 23, pp. 548–575, by Ivo Mavrovic, Consultant and A. Ray Shirley, Jr., Applied Chemical Technology.

1. E. Blasiak, *Technology of Nitrogen Compounds*, Vol. 2, State Technical Publisher, Warsaw, 1956, pp. 596–642.
2. E. Janecke, *Z. Elektrochem.* **36**, 645 (1930).
3. A. Seidell, *Solubility of Inorganic Compounds*, 2nd ed., D. Van Nostrand Co., New York.
4. D. F. Du Toit, *Proc. Koninkl. Akad. Wetenschap, Amsterdam* **16**, 555 (1913).
5. R. Escales and H. Kopke, *Chem. Zig.* **33**, 595 (1911).
6. **U.S. Pat. 3,232,984** (Feb. 1, 1966), J. A. Finneran (to Pullman, Inc.).
7. **U. S. Pat. 3,255,246** (June 7, 1966), I. M. Singer, Jr. (to E. I. du Pont de Nemours & Co., Inc.).
8. Knop and Hufner in A. E. Werner, ed., *Chemistry of Urea*, Longmans, Green and Co., London, 1923, p. 161.
9. C. Matignon and M. Friïjacques, *Bull. Soc. Chim. France* **29**, 21 (1921); **31**, 394 (1922).
10. T. R. Briggs and V. Migrdichian, *J. Phys. Chem.* **28**, 1121 (1924).
11. E. Briner, *J. Chim. Phys.* **4**, 267 (1906).
12. N. F. I. Isambert, *Compt. Rend.* **92**, 919 (1881).
13. F. Fichter and B. Becker, *Ber.* **44**, 3470 (1911).
14. C. Matignon and M. Friïjacques, *Bull. Soc. Chim.* **31**, 394 (1922).
15. K. K. Clark, V. L. Gaddy, and C. E. Rist, *Ind. Eng. Chem.* **25**, 1092 (1933).
16. *Eur. Chem. News* (Jan. 17, 1969).
17. S. Kawasumi, *Bull. Chem. Soc. Jpn.* **24**, 148; **25**, 227; **26**, 218; **27**.
18. J. D. Bridges, *TVA Bulletin Y-150*, 1980.
19. *Inorganic Chemicals Series M28A*, Bureau of the Census, U.S. Department of Commerce, Washington, D.C., annual reports, 1978–1980.
20. *Chem. Eng. News* **60**, 22 (1982).
21. A. R. Shirley, Jr., L. M. Nunnelly, and F. T. Carney, Jr., *paper presented at 182nd National Meeting of American Chemical Society, TVA Circular Z-125*, 1981.
22. *World Fertilizer Market Information Services*, TVA, National Fertilizer Development Center, Muscle Shoals, Ala.

General References

J. Meesen, "Urea" in *Ullmann's Encyclopedia of Industrial Chemistry*, 5th ed., vol. A27, VCH Publishers, 1996, Chapt., 1–7.
Jonckers, V. Mennen, J. Meesen, and W. Lemmen, *Stamicarbon New Process Urea 2000 and Stamicarbon, Wastewater Process*, Stamicarbon AB Geleen, The Netherlands, Feb. 1997.

Ivo Mavrovic
Consultant

A. Ray Shirley, Jr.
Applied Chemical Technology

G. R. "Buck" Coleman
Consultant

WATER-SOLUBLE POLYMERS

Water-soluble polymers find application in a wide variety of areas that include polymers as food sources, plasma substitutes, and as diluents in medical prescriptions. Other areas of importance for water-soluble polymers include detergents, cosmetics, sewage treatment, stabilizing agents in the production of commodity plastics, rheology modifiers in the various processes for petroleum, textile, paper, and latex coatings production. In this article the general features of polymers that influence their water solubility and solution properties will be discussed, beginning with a consideration of carbohydrate polymers provided by natural processes. Carbohydrates exhibit a broad spectrum of properties arising from structural variations of essentially one type of monomer. The discussion of synthetic water-soluble polymers will emphasize differences influenced by variation in the chemical composition of the polymer and the mechanism of polymerization. This article is intended to be a synopsis of water-soluble polymers that have significant commercial impact. A more detailed discussion of water-soluble polymers is available (1–4).

Hydrophilic Groups. Water solubility can be achieved through hydrophilic units in the backbone of a polymer, such as O and N atoms that supply lonepair electrons for hydrogen bonding to water. Solubility in water is also achieved with hydrophilic side groups (eg, OH, NH₂, CO₂, SO₃).

These groups are found in the various polymers discussed here. Truly unique in its ability to interact and promote water solubility is the $\text{--O--CH}_2\text{--CH}_2\text{--}$ group. The interactions of these groups with water and their placement in the polymer structure influence the water solubility of the polymer and its hydrodynamic volume.

Viscosity Efficiency. A majority of the applications of water-soluble polymers revolve around their role in increasing the viscosity of water solutions. The hydrodynamic volume of the polymer influences its viscosity efficiency, and thereby its ability to modify the rheology of an application formulation. The primary parameter influencing the hydrodynamic volume is the polymer's molecular weight. The second is in the conformational rigidity or extension of the polymer chain in solution. For example, carbohydrate polymers contain repeating ring structures that facilitate a more rigid structure than observed in nonionic synthetic polymers. At a given molecular weight this effects a greater viscosity efficiency. The rigidity also can be increased by inclusion of charged groups in both synthetic and carbohydrate polymers. Such groups lead to electrostatic repulsions and extension of the chain in deionized aqueous solutions and an increase in the hydrodynamic volume of the polymer. Greater rigidity also can be achieved in carbohydrate polymers if helical conformations are realized. This occurs when certain interunit positional bonding patterns are coupled with pyranosyl ring branches to the main chain.

Carbohydrate Polymers

An anhydroglucose ring or glucopyranosyl unit (Fig. 1a) is the structural unit on which human metabolism is dependent. The anhydroglucose unit provides, through the four hydroxyl units, a diversity of polymer structures that can be formed through variations in positional bonding between the rings (ie, 1 → 2, 1 → 3, 1 → 4, and 1 → 6) (5). Examples of different positional bonding in carbohydrate polymers are illustrated in Figure 2.

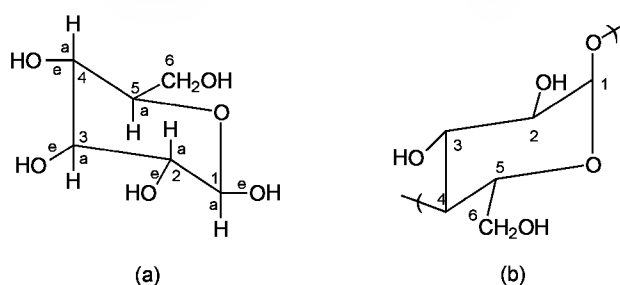


Fig. 1. (a) The glucopyranosyl basic unit, where e = equatorial and a = axial bond; (b) amylose with equatorial–axial interunit bonding (the C–H axial bonds have been omitted).

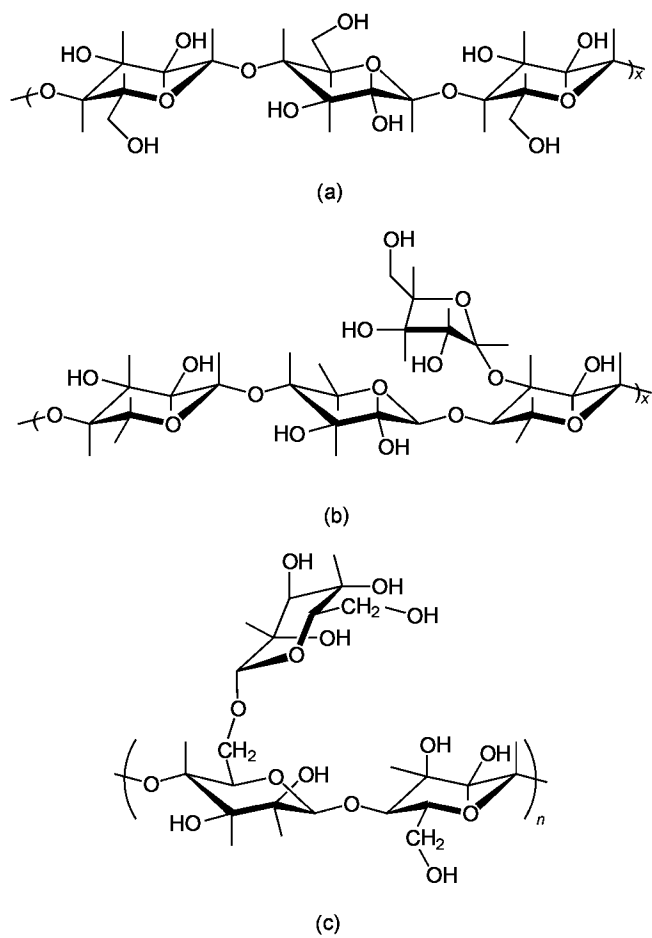


Fig. 2a. Examples of carbohydrate polymers with interunit and branch position differences: (a) cellulose; (b) flaxseed gum; (c) guaran; (d) C-6 hydroxyl; and (e) *Sclerotium glaucanum* polysaccharide (SGPS).

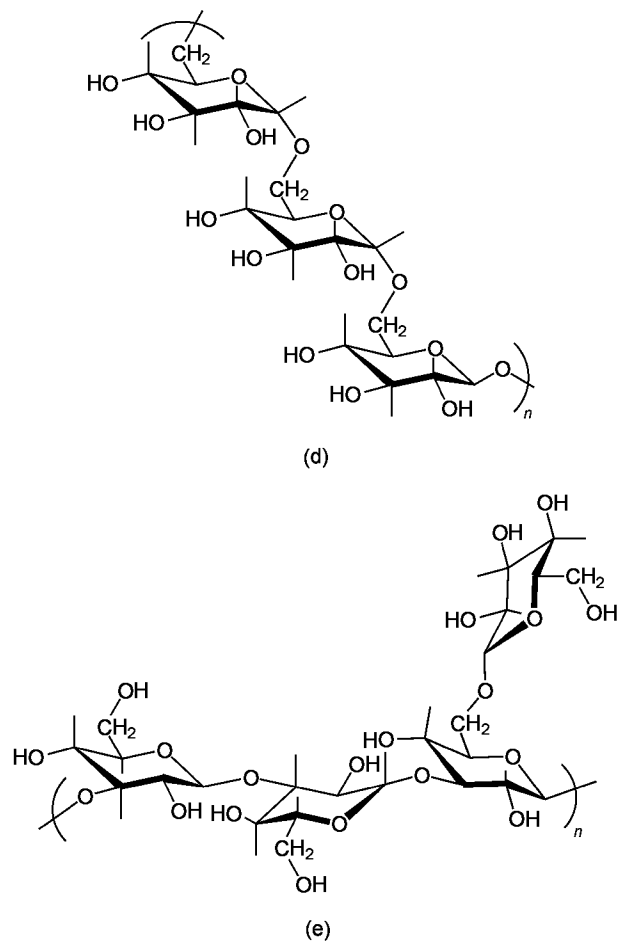


Fig. 2b. Continued

Symmetrical 1 $\rightarrow$ 4 bonding provides the world's most abundant polymer, cellulose (Fig. 2a). The bonds linking hydroxyl units in the plane of the puckered ring are referred to as equatorial bonds. The glucopyranosyl unit is also present in amylose (Fig. 1b). The variance between cellulose and amylose (the latter is a component of starch (qv)) is in the interunit bonding between anhydroglucose rings. In cellulose, the rings are connected through equatorial-equatorial bonding (beta-linkage); in amylose the 1 $\rightarrow$ 4 inter-ring bonding is equatorial-axial (alpha-linkage; axial denotes a bond perpendicular

to the general plane of the ring). The beta-linkage in cellulose (qv), complemented by the intrahydrogen bonding among rings facilitates a linear projection of the polymer. This, complemented by interhydrogen bonding among polymer chains, provides crystallinity and a rigid structure. Thus, the beta-linkage serves as a structural unit in plants and higher species. To transform the world's most abundant polymer, cellulose, into a water-soluble species requires replacement of some of the hydroxyls to disrupt the extensive hydrogen bonding.

Amylose (Fig. 1b), with alpha-interconnecting linkages, is soluble in hot water (unlike cellulose), but it retrogrades in low temperature aqueous solutions into a helical conformation and precipitates. Glucopyranose branching from the C-6 hydroxyl provides solubility at low temperatures. Thus this anomeric bonding difference (ie, alpha-(1 → 4) instead of the beta (1 → 4) linkage in cellulose) provides a readily assessible source of energy and is designated at a storage source in nature. The principal starch product from potato, corn, and other vegetable sources is the branched derivative of amylose (to inhibit precipitation at lower temperatures) known as amylopectin. In higher animal forms it is stored as a more branched derivative (ie, glycogen) to minimize solution viscosities.

The anomeric difference (ie, the alpha- and beta-linkages) between cellulose and amylose is important. Of greater importance in determining the aqueous solution properties of water-soluble polymers is the interunit bonding patterns between rings. When the bonding is between the 1 → 3 positions of the repeating glucose units (rather than 1 → 4 as in cellulose), the polymer is again water insoluble, but the addition of caustic swells the beta-(1 → 3) linked polymer. As with amylose, branches from the C-6 position promote solubility in water. *Sclerotium glaucum* polysaccharide (SGPS) is 1 → 3 linked and has 1 → 6 side-chains as shown in Figure 2e. The 1 → 3 bonding with a branch on the C-6 position of every third unit facilitates a triple helix conformation (6,7) and high thickening efficiency. Two naturally occurring carbohydrate polymers with structural main chain similarity to cellulose (ie, in the beta-(1 → 4) linkage), but with water solubility without derivatization due to side branch units, are the nonionic fraction of flaxseed gum and guaran (Fig. 2b and 2c, respectively). In Figure 2c the repeating unit is a C-2 epimer of glucopyranose (ie, the C-2 hydroxyl is axial rather than equatorial) and the main chain repeating unit is referred to as mannopyranose. The polymer in Figure 2c is known as guaran or guar gum. The cis stereochemistry in guar gum provides a source of cross-linking with borates and chromium ions that is extensively used in fracturing fluids in petroleum recovery applications (8).

When the interbonding position between repeating rings is through the C-6 hydroxyl (Fig. 2d), there is an additional methylene group between rings. This promotes an extra degree of freedom (9) and water solubility, without derivatization. The primary product with this bonding is known as dextran. The flexibility afforded with this interunit bonding also provides less thickening efficiency than the more segmentally rigid polymers containing interunit bonding through the alicyclic secondary ring hydroxyls. In medical applications in which dextrans are used this is an advantage. In most other application areas this would be a disadvantage. For a more detailed discussion of carbohydrate polymers see CARBOHYDRATES, CELLULOSE, AND STARCH (10).

CELLULOSE

Commercial Derivatization of Cellulose. Cellulose, the world's most abundant polymer, is derivatized for use in a variety of markets. Two important classes are cellulose esters (qv) and cellulose ethers (qv). Cellulose esters are not water soluble and are not discussed here; cellulose ethers are an important segment of water-soluble polymers.

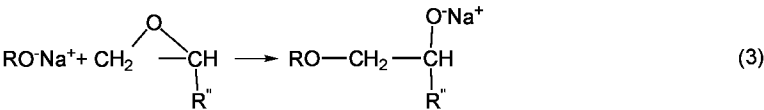
Commercial derivatization of cellulose begins with the addition of sodium hydroxide to form alkali cellulose (AC) (eq. 1, R = carbohydrate).



The AC may react with methyl chloride or alpha-chloroacetic acid via a direct displacement reaction (eq. 2). The derivatives would be methyl cellulose or carboxymethyl cellulose.



Alternatively, the AC may react with oxiranes (eg, ethylene oxide (R'' = H) or propylene oxide (R'' = CH₃); (eq. 3)); this is a catalyzed addition and requires a much lower caustic-to-cellulose ratio than is used in direct displacement (eq. 2).



The derivatives are hydroxyethyl and hydroxypropyl cellulose. All four derivatives find numerous applications and there are other reactants that can be added to cellulose, including the mixed addition of reactants leading to adducts of commercial significance. In the commercial production of mixed ethers there are economic factors to consider that include the efficiency of adduct additions (ca 40%), waste product disposal, and the method of product recovery and drying on a commercial scale. The products produced by equation 2 require heat and produce NaCl, a corrosive by-product, with each mole of adduct added. These products are produced by a paste process and require corrosion-resistant production units. The oxirane additions (eq. 3) are exothermic, and with the explosive nature of the oxiranes, require a dispersion diluent in their synthesis (see CELLULOSE ETHERS).

Biosynthesis. Although cellulose can be derivatized, such materials do not provide the optimum properties desired in many applications, such as retention of viscosity at higher solution temperatures, greater mechanical stability, and greater thickening efficiency. These properties can be approached with carbohydrate polymers in helical conformations, which can be achieved in some carbohydrate polymers prepared by fermentation processes (eg, SGPS; see Fig. 2e) prepared by a yeast (qv) fermentation process. Yeasts have an advantage over fermentation processes using bacteria, because the fermentation can be conducted in low pH media. Yeast are also larger and easier to remove by filtration (qv). However, the most successfully commercial fermentation polymer is XCPS, synthesized by a bacteria, *Xanthomonas campestris*. In this polymer the main chain is simply cellulose, but with three pyranosyl rings branched from the C-3 position (Fig. 3) of every other repeating ring. This arrangement promotes a helical conformation.

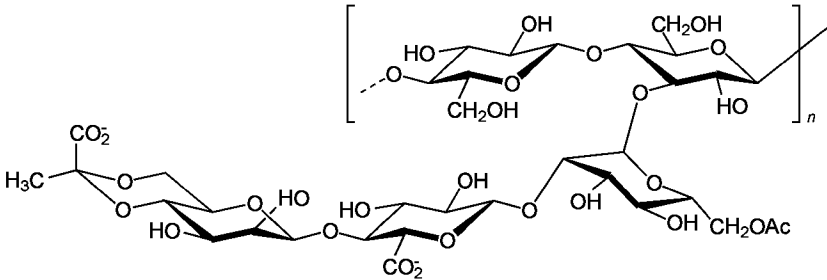


Fig. 3. Structure of *Xanthomonas campestris* polysaccharide.

With the proper ratio of nutrients and oxygen feed, a water-soluble polymer is produced and accompanied by growth in the microorganism population. Both contribute to the viscosity of the medium and this limits the production process. Fermentation processes require more strenuous mixing and control conditions.

In the carbohydrate polymers discussed so far there is variation in positional or anomeric bonding between different types of polymers, but not

within a given polymer chain. Of course such variations would provide different isomers and unique carbohydrate polymer properties. For example, the polysaccharide produced by a yeast *Pullularia pullulans* (Fig. 4) is different from amylose amylopectin products; this polymer developed by the Japanese is an effective barrier polymer (qv) (11).

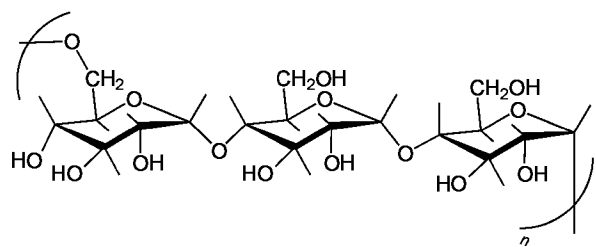


Fig. 4. Structure of *Pullularia pullulans* polysaccharide.

Biosynthesis studies of carbohydrate polymers were numerous over the decades that followed the synthesis of SGPS and XCPS; however, these studies did not lead to polymers with the molecular weight and helical characteristics necessary to achieve the optimum properties of SGPS and XCPS. More recently, two fermentation polymers have produced optimum properties through variations in positional and interunit bonding patterns. The structures of two such products (12–15), gellan and wellan, are illustrated in Figures 5 and 6.

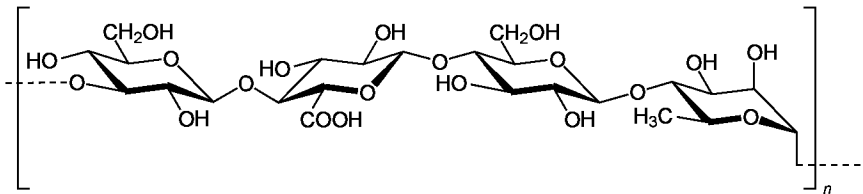


Fig. 5. Primary structure of gellan gum.

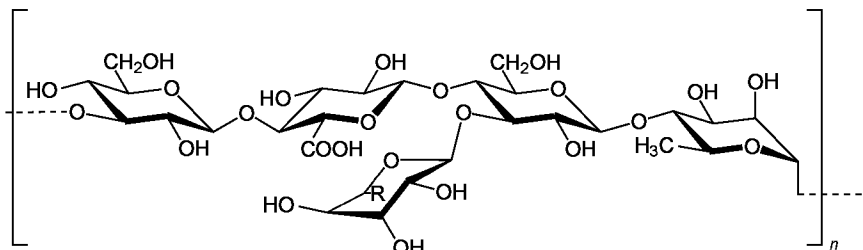


Fig. 6. Primary structure of wellan gum.

There has been growing activity in the biomodification of existing carbohydrate polymers, and although these types of studies may be too impractical to promote commercial activity in the near future, they are contributing to an understanding of structure property relationships in aqueous media (16).

Other Considerations. With multiple hydroxyl units on every repeating ring, most commercial carbohydrates particles are surface treated with glyoxal. After adequate dispersion of the particles in water, a small quantity of base solution is added to remove the acetal cross-linked structure. The individual particles then readily hydrate without the agglomeration of partially hydrated particles before complete hydration is achieved. The segmentally rigid and conformationally rigid (in helical polymers) ring structures provide a more viscous aqueous shear viscosity, but lower extensional solution viscosity for a given molecular weight (Fig. 7). The rigidity provides greater mechanical stability relative to synthetic polymers discussed below. Helix formation improves the thermal oxidation and hydrolytic stability of carbohydrate polymers related to their repeating acetal linkages, but in general they are more unstable than many of the synthetic water-soluble polymers to be discussed below.

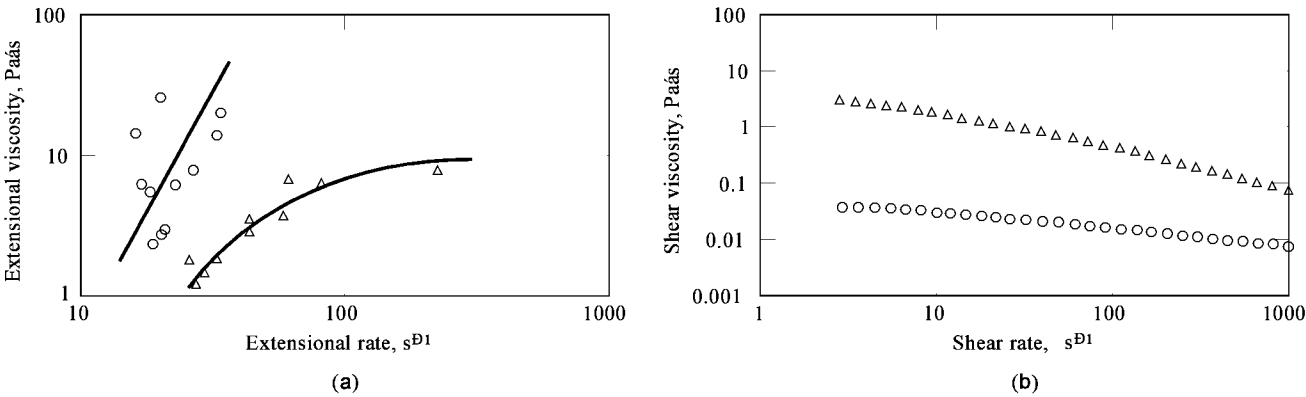
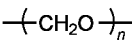


Fig. 7. (a) The dynamic uniaxial extensional viscosity (43) and (b) shear viscosity dependence of deformation rate of aqueous thickener solutions: (○), 0.3 wt % polyoxyethylene (POE) ($M_v = 6.0 \times 10^6$), (Δ), 1.0 wt % hydroxyl ethyl cellulose (HEC) ($M_v = 9.5 \times 10^5$). To convert Pa·s to P, multiply by 10.

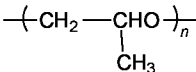
Synthetic Water-Soluble Polymers

Polyoxyethylene. Synthetic polymers with a variety of compositionally similar chemical structures are as follows. Based on polarity, poly(oxyethylene) (1) would be expected to be water soluble. It is a highly crystalline polymer used in engineering plastics, but it is not water-soluble (see

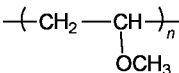
ACETAL RESINS). Polyoxypropylene (PPO) (2) and poly(methyl vinyl ether) (PMVE) (3) have



(1)

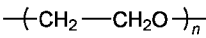


(2)



(3)

identical chemical compositions; PMVE is water soluble up to a modest temperature (~50°C), but PPO is not soluble in water. PPO is soluble only in the oligomeric form with less than 10 PO units. The hydrophobicity of the PO units overcome the terminal hydroxyl groups facilitating water solubility when there are more than 10 repeating units. In view of the lack of water solubility of two compositionally similar polymers, the water solubility of polyoxyethylene POE (4) is notable. It occurs because of the unique interaction of water with the



(4)

oxyethylene chain (17–19). Anions are effective in disrupting the interaction between the oxyethylene units and water, particularly with increasing valence of the anion (20).

Polyoxyethylene (POE) is synthesized employing different catalyst systems depending on the desired mol wt. For molecular weight below 20,000 (generally referred to as poly(ethylene glycol)s, PEGs), base or the Na⁺ or K⁺ alkoxides of methanol or butanol are used. Because of the frequent transfer reactions in this step-growth synthesis in aqueous media, a transition metal such as Cu with nitrogen ligands in a nonpolar solvent must be used to synthesize POE molecular weights in the range of 10⁷, needed for superior drag reduction properties. Without the numerous hydroxyls present in carbohydrate polymers, POE particle structures cannot be treated to minimize their hydration prior to dissolution. The particles surfaces are covered with silica to minimize hydration and blocking of POE particles on storage. POE and PEGs of various molecular weights have found numerous applications (17). The most recent interest has been in their use as anticoagulants through an effect on blood platelets (21,22). There have also been extensive studies on and commercial use of ethylene oxide–propylene oxide block copolymers. The largest single commercial growth area for PEGs in the 1990s has been their surfactant modification, one example of which is hydrophobically modified ethoxylated polyurethanes (HEUR); these are discussed later in this article.

Commodity Chain-Growth Polymers. Two of the largest commodity water-soluble polymers are poly(vinyl alcohol) (PVA) and polyacrylamide (PAM). They are prepared by the free-radical initiation of vinyl monomers, a chain-growth polymerization technique. Free-radical species involved in the propagation step terminate by coupling or disproportionation, limiting the molecular weight obtainable. The free-radical concentration must be kept below 10⁻⁸ M to obtain high molecular weights. The parameters that influence the molecular weight or degree of polymerization X_n in a free-radical polymerization process are given in equation 4,

$$X_n = \frac{k_p^2 [M]^2}{k_t R_p} \tag{4}$$

where [M] is the monomer concentration. The rate of polymerization R_p can be kept low by using low free-radical initiator concentrations. The remaining parameters are determined by the monomer. A listing of the propagation k_p and termination k_t rate constants of the propagating radicals and the chain-transfer constant of a series of water-soluble monomers is found in Table 1. From these values it is apparent that polyacrylamide (PAM) of high molecular weights can be synthesized with high purity monomer. This is not possible for poly(vinyl alcohol) (PVA). PVA is prepared by hydrolysis of vinyl acetate. Neither PVA or PAM finds extensive commercial use as a water-soluble polymer because both exhibit limited water solubility, unless a copolymer unit is randomly placed among the alcohol or amide units. This is accomplished in PVA by retaining 2 to 28% of the acetate units in the hydrolysis step (5). In PAM the comonomer unit is acrylic acid obtained through copolymerization or through hydrolysis of 10 to 50% of the amide groups to carboxylate units (6). The acid groups in their ionized forms serve to decrease the adsorption of PAMs, of economic importance in petroleum recovery applications (23).

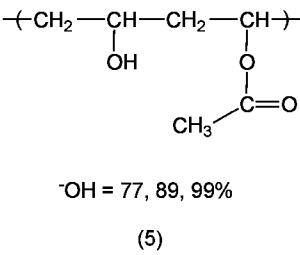
Table 1. Polymerization and Chain-Transfer Constants for Various Monomers^a

Monomer	k_p	$k_t \times 10^{-6}$	k_p/k_t	$C_m \times 10^4$
acrylamide	17,200	16.3	4.26	0.2
vinyl acetate	1,100	80	0.234	1.45
2-vinyl pyridine	186	33	0.0323	0
4-vinyl pyridine	12	3	0.00693	6.7
styrene	44	47.5	0.00638	0.28
methyl methacrylate	21	12.3	0.00599	0.26

^a k_p = propagation rate constant, k_t = termination rate constant, and C_m = chain-transfer constant. All values at 25°C.

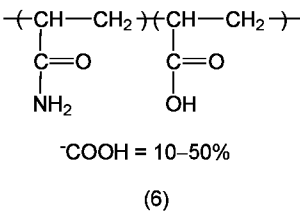
Poly(vinyl alcohol). The vinyl alcohol monomer is unstable and isomerizes to acetaldehyde. The polymer is obtained by the hydrolysis of poly(vinyl acetate). The vinyl acetate monomer produces a very high energy radical during chain-growth propagation. This accounts for the high chain transfer to monomer (see Table 1), a higher than normal head to head addition of propagating species, and grafting of some of the propagating species to polymer that is formed during an earlier stage of the polymerization. This leads to a more complex variation in structure than observed in PAM polymers, and this, with the differences that are realized with different methods of hydrolysis, can result in different Poly(vinyl alcohol) (PVA) (24,25). These factors, on hydrolysis, lead to PVA below 50,000 molecular weights. Descriptions of the industrial methods used in the synthesis of poly(vinyl acetate) polymer (26,27), and the conversion to poly(vinyl acetate) to poly(vinyl alcohol) (PVA) have been discussed both from the chemical and production aspects (28–30) (see VINYL POLYMERS, VINYL ACETATAL POLYMERS; VINYL ALCOHOL POLYMERS).

PVA is isomeric with POE; however, it can enter into hydrogen bonds with both water and with the other hydroxyl units of the repeating polymer chain, forming both inter- and intrahydrogen bonds. The extensive hydrogen bonding can lead to crystallinity, an occurrence that complicates its water solubility. Commercial PVA is essentially atactic. With most chain-growth polymers, crystallinity is associated with stereoregularity, but the small size of the hydroxyl substituent group promotes crystallinity even in the atactic polymer. For this reason poly(vinyl acetate) is seldom hydrolyzed completely; it is manufactured retaining three acetate levels: 25, 12, and 2 mole percents (these numbers represent averages). With higher acetate percentages the polymer is more surface



active, which is important in its role as a suspending agent in poly(vinyl chloride) commodity resin production. Also, it is readily soluble in water at ambient temperatures. With low acetate percentages, PVA is difficult to dissolve unless the water temperature is high, and on cooling the low acetate PVA may precipitate.

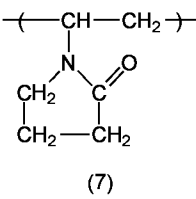
Hydrolyzed Polyacrylamide. HPAM (6) can be prepared by a free-radical process in which acrylamide is copolymerized with incremental amounts of acrylic acid or through homopolymerization of acrylamide followed by hydrolysis of some of the amide groups to carboxylate units.



The carboxylated units, ionized, decrease adsorption on subterranean substrates (23), in proportion to the number of units, an important parameter in petroleum recovery processes. In waste treatment processes cationic acrylamide comonomer units are often used (31) to increase adsorption and thereby flocculation of solids in wastewater (see *ACRYLAMIDE POLYMERS; FLOCCULATING AGENTS*). The favorable k_p , k_t and C_m characteristics of acrylamide facilitate the production of very high molecular weights. The kinetic parameters are dependent on the pH of the aqueous media and salt effects. If HPAM is obtained by copolymerization with acrylic acid, there are strong salt effects, typical of ionogen polymerizations (32).

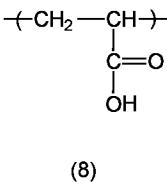
HPAM, like POE, exhibits an aging process in which the solution viscosity continues to change. This is most likely due to incomplete hydration of the dispersed polymer particle. Because the synthesis of HPAM can be conducted in water, the problem of dissolution in many applications is addressed by polymerizing the monomer in water-in-oil emulsions; aerosol OT, a branched anionic surfactant, is the primary stabilizer. The thermal stability of acrylamide solutions are generally higher than most other water-soluble polymers; however, the high molecular weight species, like their POE equivalents, have very high extensional viscosities (see Fig. 7), favorable for good drag reduction behavior. This also, however, reflects on their rapid degradation under mechanical deformation. The other negative aspect of PAM or HPAM aqueous solutions is hydrolysis of the amide group. This is most apparent in caustic flooding in secondary petroleum recovery applications. Studies have shown that the hydrolysis to carboxylate groups is limited by the electrostatic repulsion of charged carboxylate groups, below 50 mole carboxylate units there is little hydrolysis (33) of the remaining amide units.

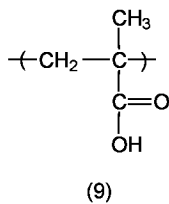
Poly(vinyl pyrrolidone). Another commercial polymer with significant usage is PVP (7). It was developed in World War II as a plasma substitute for blood



(34). This monomer polymerizes faster in 50° o water than it does in bulk (35), an abnormality inconsistent with general polymerization kinetics. This may be due to a complex with water that activates the monomer; it may also be related to the impurities in the monomer (eg, acetaldehyde, 1-methyl pyrrolidone, and 2-pyrrolidone) that are difficult to remove and that would be diluted and partitioned in a 50° o aqueous media (see *VINYL POLYMERS, N-VINYLAMIDE POLYMERS*).

Poly(acrylic acid) and Poly(methacrylic acid). Poly(acrylic acid) (8) (PAA) may be prepared by polymerization of the monomer with conventional free-radical initiators using the monomer either undiluted (36) (with cross-linker for superadsorber applications) or in aqueous solution. Photochemical polymerization (sensitized by benzoin) of methyl acrylate in ethanol solution at 78° C provides a syndiotactic form (37) that can be hydrolyzed to syndiotactic PAA. From academic studies, alkaline hydrolysis of the methyl ester requires a lower time than acid hydrolysis of the polymeric ester, and can lead to oxidative degradation of the polymer (38). Poly(methacrylic acid) (PMAA) (9) is prepared only by the direct polymerization of the acid monomer; it is not readily obtained by the hydrolysis of methyl methacrylate.

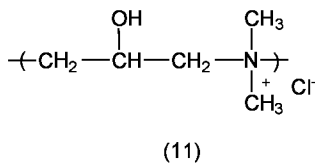
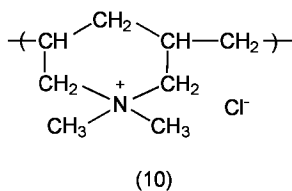




Free-radical polymerization of the monomer is a commonly used route to prepare PMAA. Over the years this has been done for both PAA and PMAA in aqueous solution, using peroxy initiators such as hydrogen peroxide or persulfate ions. Hydrogen peroxide has the advantage over other initiators in that it does not leave any organic or ionic impurities in the system. Redox systems also have been used, such as persulfate or thiosulfate and ferrous ions plus hydrogen peroxide (39,40); however, the latter system has the disadvantage of leading to an appreciable iron content for the final polymer, which can lead to polymer degradation during applications (41). Because of this the movement has been toward the use of water-soluble organic free-radical initiators to avoid the transition metals present in application formulations. Both PAA and PMAA have multiple pK_s . This property has facilitated many application uses such as dispersing agents for pigments. Carboxylate groups have strong chelating tendencies with cations, particularly divalent ones. This may lead to precipitation of polyelectrolytes, such as PAA and PMAA. To take advantage of this while maintaining solubility, maleic anhydride is copolymerized with styrene. The copolymerization involves an electron-donor acceptor complex that produces an alternating copolymer (42). The anhydride units are hydrolyzed to vicinal carboxylates which enhances the chelating ability over 1,3-carboxylate units, and the aromatic units are sulfonated to maintain solubility in high salinity environments. This type of product has found utility in boiler applications. Unfortunately, the degree of sulfonation of the aromatic unit has not been above 10% and this has limited its market potential.

CATIONIC WATER-SOLUBLE POLYMERS

Cationic monomers are used to enhance adsorption on waste solids and facilitate flocculation (31). One of the first used in water treatment processes (10) is obtained by the cyclization of dimethyldiallylammonium chloride in 60–70 wt % aqueous solution (43) (see WATER). Another cationic water-soluble polymer, poly(dimethylamine-*co*-epichlorohydrin) (11), prepared by the step-growth

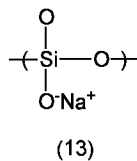
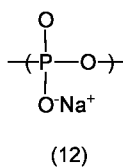


polymerization of dimethyl amine and epichlorohydrin, is also used for adsorption on clays (qv). This polymer stabilizes shale in drilling fluid formulations in petroleum applications (44). In addition to these cationic polymers used in commodity applications, cationics find use in specialty products such as cosmetic applications. An example of this is the addition of an epichlorohydrin containing an amine group to hydroxyethylcellulose. At a molar substitution of 0.2 to 0.4 per repeating glucopyranosyl ring (see Fig. 1a) the pendent amine provides substantivity to skin and hair (45) and has proven profitability comparable with Carbopol 940.

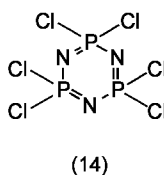
The literature on the preparation and application of numerous synthetic cationic quaternary water-soluble polymers has been reviewed in detail (46).

INORGANIC WATER-SOLUBLE POLYMERS

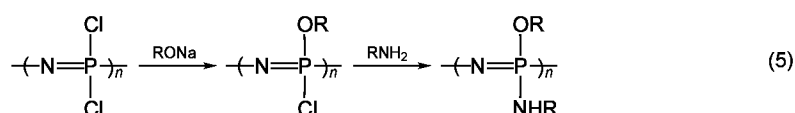
Two inorganic water-soluble polymers, both polyelectrolytes in their sodium salt forms, have been known for some time: poly(phosphoric acid) (12) and poly(silicic acid) (13). A more exciting inorganic water-soluble polymer with nonionic



characteristics has been reported (47). This family of phosphazene polymers is prepared by the ring-opening polymerization of a heterocyclic monomer (14) followed by replacement of the chlorine atoms in the resultant polymer. The



chlorine replacement step can be carried out either to introduce only one type of side group or, by simultaneous or sequential substitution, to introduce two or more different types of side groups (see INORGANIC HIGH POLYMERS).



Qualitative direct displacements on a polymer chain are rare, but possible for $n \geq 15,000$ in the example shown in equation 5, due to the high reactivity of the P–Cl bond. Some of the organic groups can be placed on the heterocyclic monomer before the ring-opening polymerization if their size is limited to avoid steric influences on the polymerization. The hydrophobic inorganic backbone thus provides an easy route to variable water-solubilizing side nonionic or ionizing groups (Fig. 8). These are stable to hydrolysis at room temperature. The glucosyl (**15**) and glyceryl (**16**) species are sensitive to hydrolysis in neutral pH water at 100°C; they hydrolyze slowly to phosphate, small amounts of ammonia, and glucose or glycerol when the pH is changed. The methylamino-substituted (**17**) and the poly(bis(methoxyethoxy)phosphazene) (MEEP) (**18**) are stable to water at neutral and basic pH but are sensitive to strong acids.

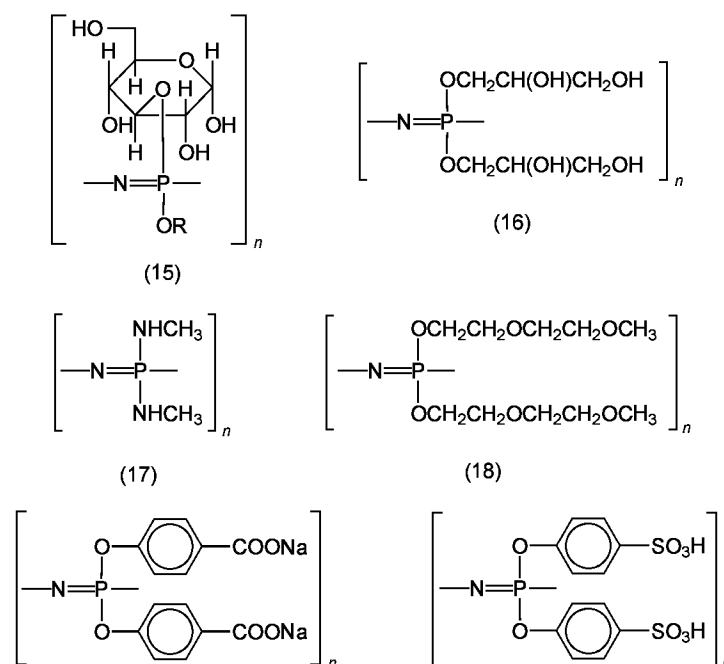


Fig. 8. Water-soluble polyphosphazenes.

One of the main advantages of water-soluble polyphosphazenes is the ease with which water-solubilizing side groups can sufficiently cross-link in a stable matrix with high energy radiation such as x-rays, gamma-rays, electron beams, or ultraviolet light. The mechanism of this reaction involves abstraction of hydrogen radicals from the methylene units contiguous to the repeating oxygen bonds in MEEP or to amine groups in other derivatives, followed by cross-combination of the resultant carbon radicals (Fig. 9). In MEEP there are 22 C-H bonds available per repeating unit.

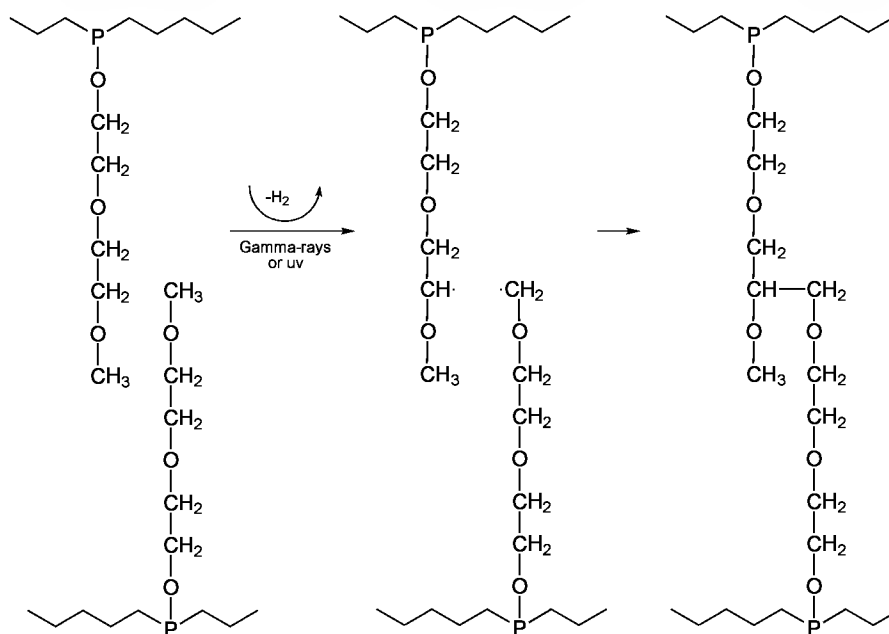


Fig. 9. Cross-linking reactions with polyphosphazenes.

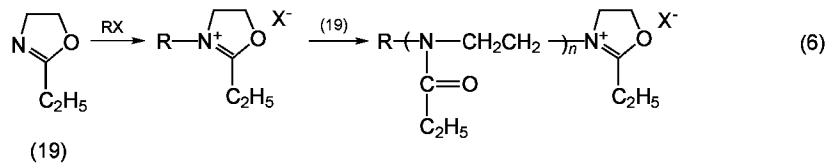
The versatility of water-soluble polyphosphazenes is in the variations in the structures that can be prepared. Structures with a low glass-transition temperature backbone can be modified with a variety of versatile side units. These may find use in solid polymeric ionic conductors, as a means to entrap and immobilize enzymes with retention of enzymic activity, and in biological functions as hydrogels with the capability of exhibiting biocompatibility and bioerodibility properties.

New Commercial Water-Soluble Polymers

The commercialization of new water-soluble polymers is most often a slow process. For example, hydroxyethylcellulose (HEC) was envisaged in 1937 by

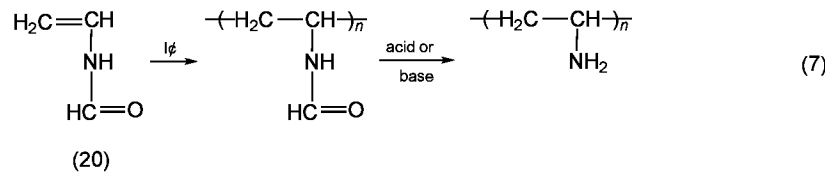
A. E. Broderick (Union Carbide). HEC did not become a viable commercial product until the early 1960s. In addition to the general production problems and market development costs, new products face a variety of environmental controls in the 1990s that add more constraints to market development. None the less two more recently developed water-soluble polymers have achieved limited market acceptance and are described below.

One product is poly(2-ethyl-2-oxazoline) (PEOX). It is prepared by the ring-opening polymerization of 2-ethyl-2-oxazoline (19) with a cationic initiator (48) (eq. 6).



Most of the polymer's characteristics stem from its molecular structure, which like POE, promotes solubility in a variety of solvents in addition to water. It exhibits Newtonian rheology and is mechanically stable relative to other thermoplastics. It also forms miscible blends with a variety of other polymers. The water solubility and hot meltable characteristics promote adhesion in a number of applications. PEOX has been observed to promote adhesion comparable with PVP and PVA on aluminum foil, cellophane, nylon, poly(methyl methacrylate), and poly(ethylene terephthalate), and in composite systems improved tensile strength and Izod impact properties have been noted.

Other fairly recent commercial products, poly(vinyl amine) and poly(vinyl amine *co*-vinyl alcohol), have addressed the need for primary amines and their selective reactivity. Prior efforts to synthesize poly(vinyl amine) have been limited because of the difficulty hydrolyzing the intermediate polymers. The current product is prepared from *N*-ethenylformamide (20) formed from the reaction of acetaldehyde and formamide. The vinyl amide is polymerized with a free-radical initiator, then hydrolyzed (eq. 7).



The protonated form of poly(vinyl amine) (PVAm-HCl) has two advantages over many cationic polymers: high cationic charge densities are possible and the pendent primary amines have high reactivity. It has been applied in water treatment, paper making, and textiles (qv). The protonated forms modified with low molecular weight aldehydes are useful as fines and filler retention agents and are in use with recycled fibers. As with all new products, unexpected applications, such as in clear antiperspirants, have been found. It is useful in many metal complexation applications (49).

Hydrophobe-Modification of Water-Soluble Polymers

Although many of the new water-soluble polymers discussed above have not achieved large-scale commercial acceptance, there is a class that has achieved outstanding success since the early 1980s: hydrophobically modified water-soluble polymers (HMWSPs). They have filled certain voids in a number of applications that include cosmetic, paper, architectural, and original equipment manufacturing (OEM) coating areas and have found unsuspected application in the airplane deicers market. The driving force for the development of HMWSPs is threefold in most application areas:

- (1) The achievement of high viscosities at low shear rates without high molecular weights, and therefore regain of high viscosities at low shear rates, when the application deformation rates are released. High mol wt polymers are susceptible to mechanical degradation at high deformation rates and permanent loss of high viscosity at low deformation rates.
- (2) Minimization of the elastic behavior of the fluid at high deformation rates that are present when high molecular weight water-soluble polymers are used to obtain cost-efficient viscosities at low shear rates.
- (3) Providing colloidal stability to disperse phases in aqueous media, not achievable with traditional water-soluble polymers.

The first criterion was associated with improved secondary and tertiary petroleum recovery processes. This is the justification for the patent applications issued to the Dow (50) and Exxon (51) corporations. The additional costs of production and the increased adsorption of such modified water-soluble polymers are detrimental to the commercial application of such polymers and even the academic studies in this area have decreased in recent years.

The second driving force is most evident in the application of coatings by roll (52) or spray (53). The higher extensional viscosity (see Fig. 7) imparted by higher molecular weight thickeners results in spatter to coatings during roll applications and inhibits misting of spray-applied coatings. Lower molecular weight associative thickeners provide low shear rate viscosities without imparting high extensional viscosities.

The third driving force for acceptance lies in the ability of hydrophobically modified, water-soluble polymers to stabilize the disperse phases of a coating formulation. For example, nonassociative thickeners (those that are not hydrophobically modified) tend to flocculate the disperse phases and this increases the viscosity at low shear rates. This is incorrectly related to the poor flow and leveling, and poor film gloss of an applied coating film. Adsorption and osmotic stabilization of the film-forming polymer colloid and the pigments in a coating formulation are the mechanisms by which the associative thickeners contribute to the improved properties of a coating, and this does not always parallel the viscosities at low shear rates.

The mechanism of adsorption onto the film-forming organic phase occurs through direct hydrophobe adsorption on the disperse phase. Aqueous polyurethane dispersions, used in OEM coatings, are prepared by a step-growth polymerization without external surfactant. Hydrophobically modified, ethoxylated urethane (HEUR) associative thickeners directly adsorb on this disperse phase and thickening is achieved without phase separation (54). In traditional latices used in architectural coatings, the type and concentration of surfactant used in the latices synthesis and added with other components in the coating compete with the hydrophobes of the thickener in the adsorption process (55). This type of competition makes associative thickeners sensitive to variations in other component changes in a coating formulation.

Adsorption onto pigments in alkaline exterior coating formulations requires that the pigment's surface be treated to encourage adsorption of associative thickeners on high energy surfaces above their isoelectric points. For example, the HEUR type of associative thickener adsorbs on TiO_2 in a coating if that surface contains a preadsorbed dispersant with hydrophobic units to associate with the hydrophobes of the thickener (56). This also appears to be the mechanism of adsorption for another class of successful associative thickener, hydrophobically modified hydroxyethylcellulose. With still another class of commercial associative thickener, hydrophobically modified, alkali-swellaible emulsions, the mechanism is different. This thickener with multiple carboxylic acid functions adsorbs directly on the pigment.

Chain-Growth Associative Thickeners. Preparation of hydrophobically modified, water-soluble polymer in aqueous media by a chain-growth mechanism presents a unique challenge in that the hydrophobically modified monomers are surface active and form micelles (50). Although the initiation and propagation occurs primarily in the aqueous phase, when the propagating radical enters the micelle the hydrophobically modified monomers then polymerize in blocks. In addition, the hydrophobically modified monomer possesses a different reactivity ratio (42) than the unmodified monomer, and the composition of the polymer chain therefore varies considerably with conversion (57). The most extensively studied monomer of this class has been acrylamide, but there have been others such as the modification of PVAlc. Pyridine (58) was one of the first chain-growth polymers to be hydrophobically modified. This modification is a post-polymerization alkylation reaction and produces a random distribution of hydrophobic units.

Hydrophobically Modified Alkali-Swellable Emulsions.

The hydrophobe modification of acrylic acid represents an important class of

hydrophobe-modified thickeners prepared by a chain-growth free-radical process. They differ slightly from other examples in that these products are generally cross-linked. One of the most successful commercial products of an acrylic acid (cross-linked variety) is Carbopol 940 (B.F. Goodrich). It possesses an efficient viscosity profile with excellent yield stress characteristics. These characteristics are associated with microviscosity control in the interior of the cross-linked matrix (59). This type of product has been hydrophobically modified for sunscreen applications in cosmetics.

Products of this type, used in textile printing and paper coating applications, have been referred to as alkali-swellable emulsions (ASE). The hydrophobe-modification of ASE resins has led to materials that have achieved a significant market penetration in the water-borne coatings area (60,61) and are designated HASE thickeners (Fig. 10). There can be a marked variation in thickener behavior with the size and number of hydrophobes and with the amount of acrylic acid in such polymers. In materials examined in one study (62), the dominant contribution to viscosity was from the electrostatic repulsions of the carboxylate groups in alkaline media. It is an interesting family of associative thickeners in which the mechanisms for thickening and gloss control in coating formulations are different from those of other hydrophobically modified, water-soluble polymers.

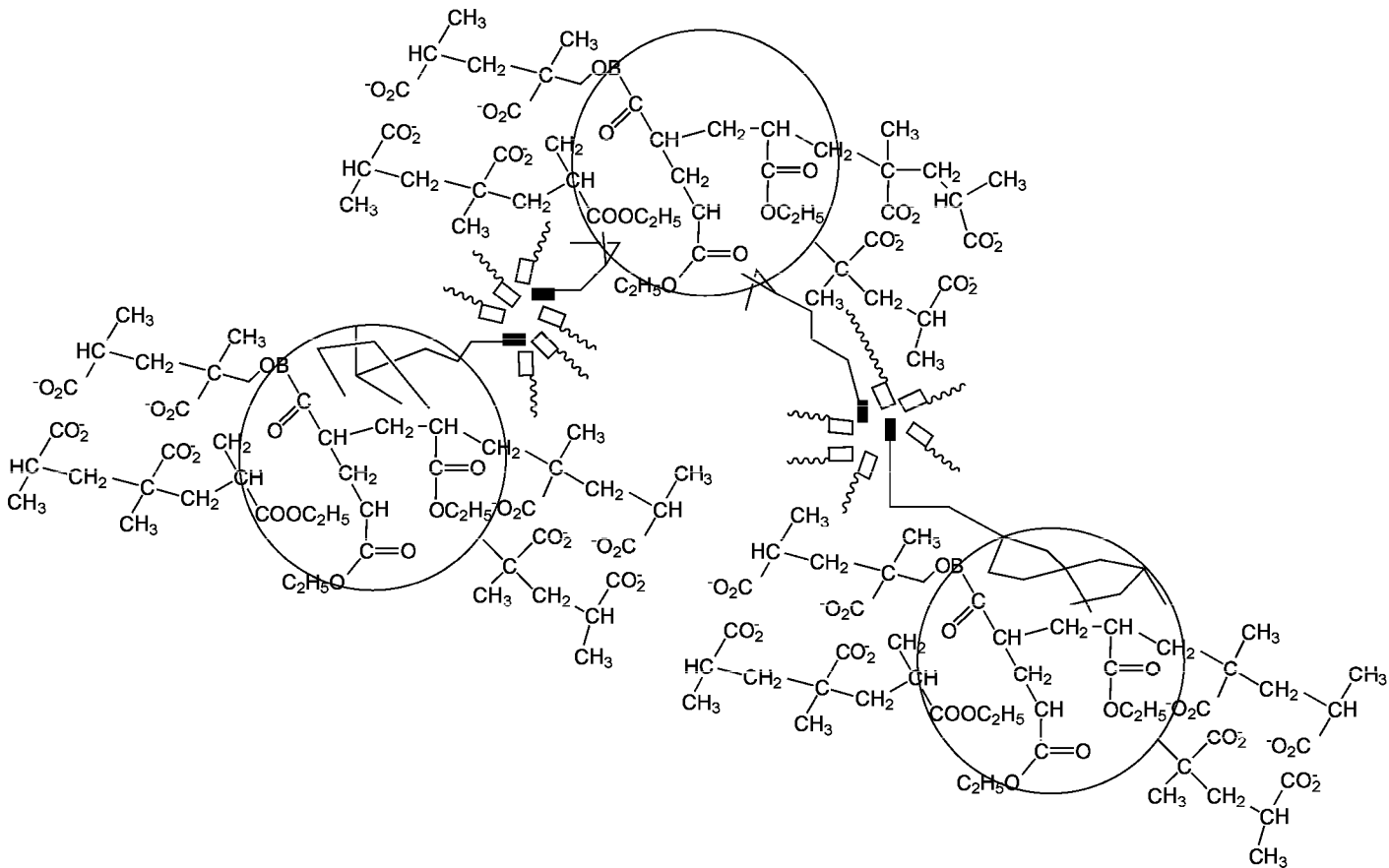


Fig. 10. General structure of a HASE thickener.

Hydrophobically Modified Carbohydrate Polymers. Hydrophobe-modification of hydroxyethylcellulose (63) produces what should be considered model associative thickeners, for the distribution of hydroxyethyl units has been characterized (64). The commercial material, a Hercules product, contains three hydrophobes per chain (65). Thus the structure of HMHEC could be represented by Figure 11. This product promotes higher applied film gloss if the pigment is treated with a hydrophobe containing dispersant. Other carbohydrate polymers such as guaran have been hydrophobically modified, but are unlikely to achieve commercial success as water-borne coating thickeners because the repeating cis vicinal diol are reactive with the borate compositions in wallboards.

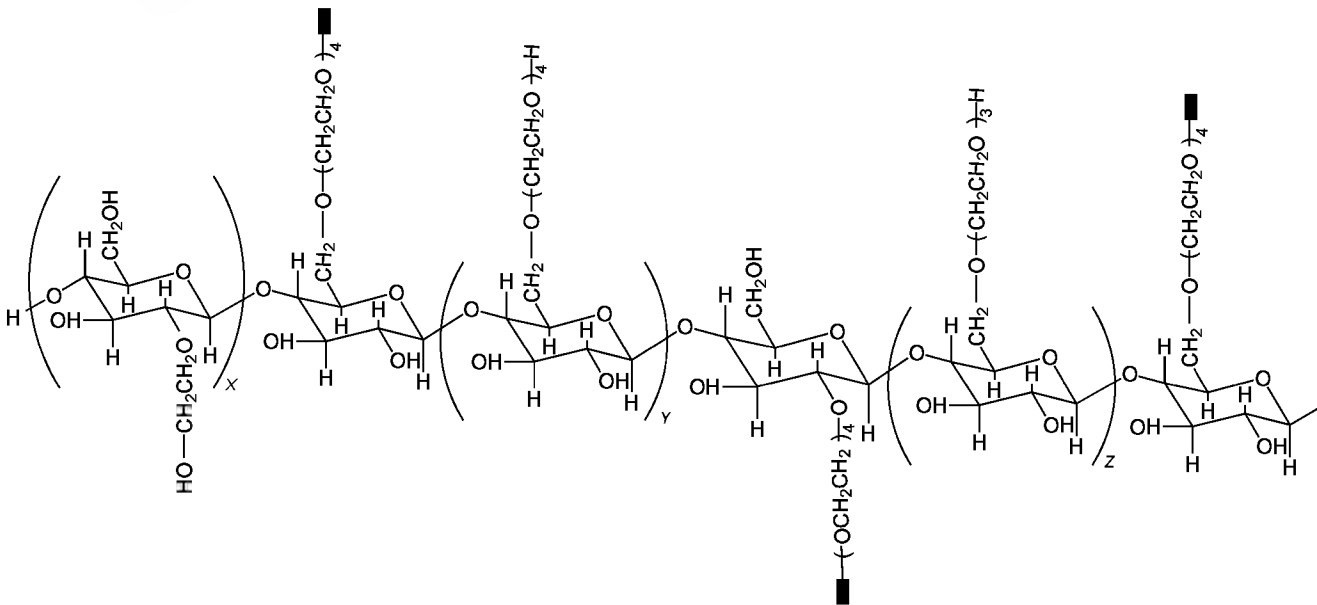
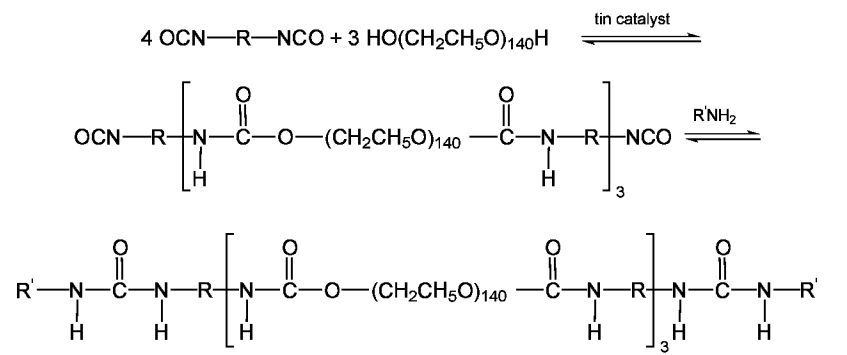


Fig. 11. General structure of a MNHEC thickener.

Hydrophobically Modified, Ethoxylated Urethane. HEUR associative thickeners are in effect poly(oxyethylene) polymers that contain terminal hydrophobe units (66). They can be synthesized via esterification with monoacids, tosylation reactions, or direct reaction with monoisocyanates. There are problems associated with all of the methods of synthesis. The general commercial procedure for their synthesis is by a step-growth addition of

6000 mol w-PEG with a slight excess of a diisocyanate to produce a telechelic adduct with terminal isocyanate groups that can then react with a hydrophobic amine or alcohol to produce a terminal hydrophobe-modified POE. This type of HEUR contains



internal hydrophobes from the reacted diisocyanate units, but these have been found ineffective in forming hydrophobic domains (67,68) and therefore in controlling the rheology of the formulated coating. Although this type of polymer seems very direct in its synthesis, the reactions of isocyanates with water and the opportunity for ready degradation of oxyethylene chains make the synthesis of hydrophobe-modified oxyethylene polymers more complex than first apparent. It is imperative that such materials be well characterized. Such thickeners, however, if prepared by a non-step growth telechelic uniHEUR process (69) offer the only real model materials for understanding the fundamental aspects of associative thickener behavior. Because these types of associative thickeners are, compositionally, mostly oxyethylene groups, they offer a greater latitude in rheology control in surfactant solutions. Those with a greater interest in these materials should see References 2–4.

It is evident that the area of water-soluble polymer covers a multitude of applications and encompasses a broad spectrum of compositions. Proteins (qv) and other biological materials are covered elsewhere in the *Encyclopedia*. One of the products of this type, poly(aspartic acid), may be developed into interesting biodegradable commercial applications (70,71).

BIBLIOGRAPHY

1.

P. Molyneux, *Water-Soluble Synthetic Polymers: Properties and Behavior*, Vols. I and II, CRC Press, Boca Raton, Fla., 1982.

2.

J. E. Glass, ed., *Water-Soluble Polymers: Beauty with Performance*, Advances in Chemistry Series 213, American Chemical Society, Washington, D.C., 1986.

3.

J. E. Glass, ed., *Polymers in Aqueous Media, Performance through Association*, Advances in Chemistry 223, American Chemical Society, Washington, D.C., 1989.

4.

J. E. Glass, ed., *Hydrophilic Polymers: Performance with Environmental Acceptance*, Advances in Chemistry Series 248, American Society, Washington, D.C., 1995.

5.

J. E. Glass, Ref. 2, Chapt. 1.

6.

Y. Deslandes, R. H. Marchssault, and A. Sarko, *Macromol.* **13**, 1466 (1980).

7.

T. Itou, T. Teramoto, M. Makasuke, and H. Suga, *Macromol.* **19**, 1234 (1986).

8.

R. K. Prud'homme, Ref. 3, Chapt. 6.

9.

K. Matsuo, *Macromol.* **17**, 449 (1984).

10.

G. O. Aspinall, ed., *The Polysaccharides*, Vols. 1, 2, and 3, Academic Press, Inc., New York, 1982.

11.

S. Yuen, *Process Biochem.* **9**, 7–9 (1974).

12.

R. Moorhouse, in M. Yalpani, ed., *Industrial Polysaccharides: Genetic Engineering, Structure/Property Relations and Applications*, Elsevier Science Publishers BV, Amsterdam, the Netherlands, 1987.

13.

T. A. Talashek and D. A. Bryant, *Carbohydr. Res.*, 303 (1987).

14.

J. W. Powell, C. F. Parks, and J. M. Scheult, *Soc. Petrol. Eng.* (5339) (Apr. 1975).

15.

K. S. Kang and D. J. Pettitt, *Industrial Gums*, Academic Press, Inc., New York, 1993.

16.

"Biopolymers as Advanced Materials", *Spring National ACS Meeting*, Anaheim, Calif.; preprints, *Polym. Matl.: Sci. Eng.* **72** (1985).

17.

F. E. Bailey and J. V. Koleske, *Poly(Ethylene Oxide)*, Academic Press, Inc., New York, 1976.

18.

J. L. Koneg and A. C. Angood, *J. Polym. Sci.* **8**(A-2), 1787 (1970).

19.

M. J. Hey, S. M. Ilett, M. Mortimer, and G. Oates, *J. Chem. Soc. Faraday Trans.* **86**(14), 2673 (1990).

20.

R. D. Lundberg, F. E. Bailey, and R. W. Callard, *J. Polym. Sci.* **4**(A-1), 1563 (1966).

21.

J. D. Andrade, V. Hlady, and S.-I. Jeon, Ref. 4, Chapt. 3.

22.

J. Li, J. Carlsson, S.-C. Huang, and K. D. Caldwell, Ref. 4, Chapt. 4.

23.

D. Martin and N. S. Sherwood, *Soc. Petrol. Eng.* (5339) (Apr. 1975).

24.

R. K. Tubbs, H. K. Inskip, and P. M. Subramanian, in C. A. Finch, ed., *Properties and Applications of Polyvinyl Alcohol*, SCI Monograph No. 30, Society of Chemical Industry, London, 1968, p. 88.

25.

R. K. Tubbs, *J. Polym. Sci.* **4**(A-1), 623 (1966).

26.

H. Shohata, in Ref. 24, p. 18.

27.

K. Noro, in C. A. Finch, ed., *Polyvinyl: Properties and Applications*, John Wiley & Sons, Ltd., Chichester, U.K., 1973, Chapt. 3 and 4.

28.

E. Hackel, in Ref. 24, p. 18.

29.

J. G. Pritchard, *Poly(Vinyl Alcohol): Basic Properties and Uses, Polymer Monographs*, Vol. 4, Gordon & Breach, New York, 1970.

30.

C. A. Finch, ed., *Polyvinyl Alcohol: Properties and Application*, John Wiley & Sons, Ltd., Chichester, U.K., 1973, Chapt. 9.

31.

P. A. Rey and Varsanik, Ref. 2, Chapt. 7.

32.

T. W. Beihoffer, D. J. Lundberg, and J. E. Glass, Ref. 3, Chapt. 8.

33.

J. Klein and R. Heitzmann, *Makromol. Chem.* **179**, 1895 (1978).

34.

H. A. Ravin, A. M. Seligman, and J. Fine, *N. Engl. J. Med.* **247**, 921 (1952).

35.

E. Senogles and R. Thomas, *J. Polym. Sci. Sympos.* **49**, 203 (1975).

36.

H. L. Greenwald and L. S. Luskin, in R. L. Davidson, ed., *Handbook of Water-Soluble Gums and Resins*, McGraw-Hill Book Co., Inc., New York, 1980, Chapt. 17.

37.

P. Monjol, *C.R. Acad. Sci. Ser. C*, **265**, 1426 (1967).

38.

A. Katchalsky and H. Eisenberg, *J. Polym. Sci.* **6**, 145 (1951).

39.

A. Oth and P. Doty, *J. Phys. Chem.* **56**, 43 (1952).

40.

R. Arnold and S. R. Caplan, *Trans. Faraday Soc.* **51**, 857 (1955).

41.

J. E. Glass, A. Ahmed, D. A. Soules, S. K. England-Jongewaard, and R. H. Fernando, *Soc. Petrol. Eng.* (11691) (1983).

42.

G. Odian, *Principles of Polymerization*, 3rd ed., John Wiley & Sons, Inc., New York, 1991, Chapt. 5.

43. G. Butler, N. Z. Zhang, "Water-Soluble Polymers," in S. W. Shalaby, C. L. McCormick, and G. B. Butler, eds., *ACS Symposium*, Series 467, Chapt. 2.

44. T. W. Beihoffer, D. S. Dorrough, C. A. K. Deem, D. D. Schmidt, and R. P. Bray, *Oil Gas J.* **47** (May 16, 1992).

45. E. D. Goddard and P. S. Leung, *Langmuir*, **8**, 1499 (1992).

46. M. F. Hoover, *J. Macromol. Sci.-Chem.* **A4**, 1327 (1970).

47. H. R. Allcock, Ref. 4, Chapt. 1 and references therein.

48. T. T. Chiu, B. P. Thill, and Fairchok, Ref. 2, Chapt. 23.

49. K. Geckeler, K. Weingartner, and E. Bayer, in E. J. Goethals, ed., *Polymeric Amines and Ammonium Salts*, Pergamon Press, Oxford, U.K., 1980, p. 277.

50. W. J. Peer, Ref. 3, Chapt. 20.

51. P. L. Valint, J. Bock, D. N. Schulz, D. B. Siano, S. J. Pace, Ref. 3, Chapt. 21 and 22.

52. J. E. Glass, *J. Coatings Technol.* **50**(641), 56 (1978).

53. L. L. Xing, J. E. Glass, and R. H. Fernando, "Technology for Water-Borne Coatings," in J. E. Glass, ed., *ACS Symposium*, Series 663, American Chemical Society, Washington, D.C., 1997, Chapt. 15.

54. J. P. Kaczmariski, R. H. Fernando, and J. E. Glass, *J. Coatings Technol.* **65**(818), 39 (1993).

55. Z. Ma, M. Chen, and J. E. Glass, *Colloids and Surfaces*, **112**(2-3) (1996).

56. D. J. Lundberg and J. E. Glass, *J. Coatings Technol.* **64**(807), 53 (1992).

57. S. Biggs, A. Hill, J. Selb, and F. Candau, *J. Phys. Chem.* **96**(3), 1505 (1992).

58. U. P. Strauss, Ref. 3, Chapt. 16.

59. R. Y. Lochhead, J. Davidson, and G. M. Thomas, Ref. 3, Chapt. 7.

60. G. D. Shay, Ref. 3, Chapt. 25.

61. R. D. Jenkins, L. M. DeLong, and D. R. Bassett, Ref. 4, Chapt. 23.

62. M.-R. Tarng and J. E. Glass, *Proc. ACS Div. Polym. Materials: Sci. Eng.* **67**, 280 (1992).

63. A. C. Sau and L. M. Landol, Ref. 3, Chapt. 18.

64. J. E. Glass, A. M. Buettner, R. W. Lowther, C. S. Young, and L. A. Cosby, *Carbohydr. Res.* **84**, 245–163 (1980).

65. J. W. Goodwin, R. W. Hughes, C. K. Lam, J. A. Miles, and B. C. H. Warren, Ref. 3, Chapt. 19.

66. W. H. Wetzel, M. Chen, and J. E. Glass, Ref. 4, Chapt. 10; other chapters also discuss HEUR thickeners.

67. J. P. Kaczmariski and J. E. Glass, *Langmuir*, **10**(9), 3035 (1994).

68. G. Fonnum, J. Blakke, and F. K. Hansen, *Colloid Polym. Sci.* **271**, 380 (1993).

69. R. May, J. P. Kaczmariski, and J. E. Glass, *Macromol.* **29**, 4745 (1996).

70. K. C. Low, A. P. Wheller, and L. P. Koslan, Ref. 4, Chapt. 6.

71. S. K. Wolk, G. Swift, Y. H. Paik, K. M. Yocom, R. L. Smith, and E. S. Simon, *Macromol.* **27**, 7613 (1994).

J. Edward Glass
North Dakota State University

WATTLE BARL.

See LEATHER.

WEIGHING AND PROPORTIONING

Weighing is the operation of determining the mass of any material as represented by one or more objects or by a quantity of bulk material. Proportioning is the control, by weighing, of relative quantities of two or more ingredients according to a specific recipe in order to make a mixed product, or to prepare the ingredients for use in a chemical process.

It is likely that volumetric measures were used for quantity determination when commodities were first bartered; however, it has been established with certainty that weighing scales or balances have been in use for at least 7,000 years (1). Measuring by weight instead of by volume eliminates some very considerable inaccuracies from, for example, changes in specific gravity of liquids with temperature, or changes in density of solids owing to voids.

Mass is a fundamental physical property of matter; the mass of an object does not change with its location on the earth. Mass is what is measured in the act of weighing, and scales are calibrated to read in units of mass, eg, kilograms. The terms weight and weighing are used somewhat loosely in this connection, in that weight is a measure of the gravitational force of the earth acting on a mass. Thus, weight depends on the location of the scale relative to the equator and on its altitude. It is common practice to calibrate scales *in situ* using mass standards (test weights); hence, for practical purposes scales actually determine mass.

The terms *balance* and *scale* are currently used to describe weighing machines of various forms. The word balance comes from the Latin term *libra bilanx*. *Libra* means scale and *bilanx* means two pans; thus *libra bilanx* means scale with two pans (2). Over the centuries *bilanx* evolved into balance. Balance still refers to sensitive weighing devices with two pans (Fig. 1), although it is also used in a more general sense to refer to any very sensitive laboratory weighing device. The word scale comes from the old Norse word *skal*, which means bowl, and it was originally used to describe the pan of a balance (2). Scale today (ca 1997) is used to describe general-purpose weighing machines found in industrial and retail applications.

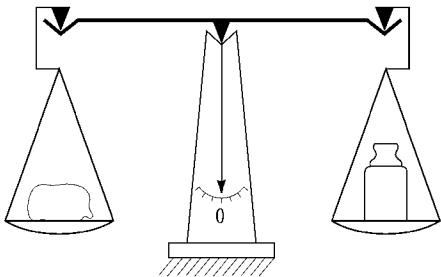


Fig. 1. Balance.

Some common industrial weighing applications include the following:

- Verifying quantities of incoming raw materials
- Controlling ingredients to the proper proportions
- Putting the product into packages of uniform weight, either by packaging directly on a scale or by using a scale to check the performance of filling equipment
- Weighing outgoing shipments for purposes of billing and determination of transportation charges
- Weighing interdepartmental material transfers for accounting purposes
- Controlling the rate of material flow to various pieces of equipment such as grinders or kilns

Weighing Principles

There are many types of scales using many different principles of operation. There are, however, three distinct elements which can be identified regardless of the principle of operation (Fig. 2).

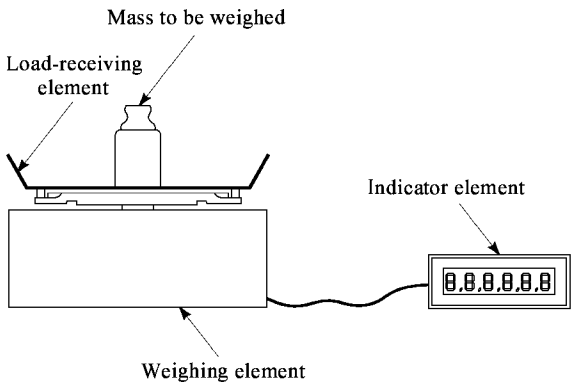


Fig. 2. Basic elements of a scale.

The load-receiving element supports the load during the weighing operation. It may take the form of a scoop, platter, deck, rail, hopper, belt

conveyor, or any other configuration appropriate to the material being weighed.

The weighing element supports the load-receiving element. It produces a signal proportional to the load and sends this signal to the indicator element. The weighing element can take many forms. It could, for example, be a mechanical lever system sending a force, a hydraulic load cell sending fluid pressure, or a strain gauge load cell sending an electrical signal.

The indicator element measures the signal from the weighing element, and converts it into a readable form. It may be any of several different types, eg, the graduated beam of Figure 3, a Bourdon tube pressure gauge, or a numeric display device. Today (ca 1997), with the increase in automation, the indicator element may not display the weight but may instead transmit it electronically to a controller.

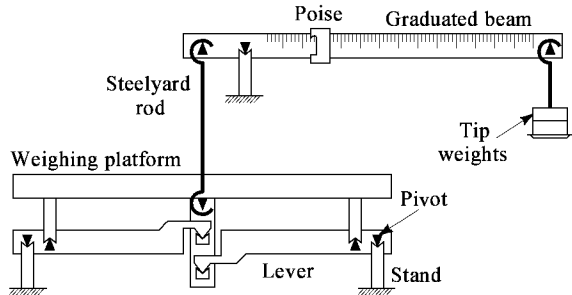


Fig. 3. Mechanical beam scale.

The principal weighing technologies in use currently are mechanical, hydraulic, strain-gauge, electromagnetic force compensation, and nuclear.

Mechanical.

Mechanical scales, measure a load either by comparing it with a known weight or by measuring the distortion of a spring caused by the load. With the even-balance type shown in Figure 1, the load is measured by direct comparison with known weights. With beam scales (Fig. 3), the force resulting from the load on the platform is reduced to a more manageable magnitude by the multiple of the lever system. It is then measured by the position of a known weight, or poise, on a graduated beam. A scale with a graduated beam and poise weight is often referred to as a Roman or steelyard scale; it was developed by the Romans and remains in limited production today (1). In operation, the poise is moved by hand until the beam balances in a horizontal position, at which point the weight can be read directly from the beam. To extend the capacity of the scale without a loss of resolution, many beam scales also use tip weights that can be placed on a hanger at the tip of the beam. For instance, a 500 kg scale could have a beam graduated from 0 to 50 kg in 0.2 kg divisions, and be supplied with a combination of tip weights which in total can counterbalance 450 kg; note, with a scale multiple of 100, a tip weight of 100 kg would actually weigh 1 kg. With this scale, any load up to 500 kg can be weighed with 0.2 kg divisions. With the poise at 50 kg, tip weights are added until the beam moves down, then the poise is moved back until the beam balances. The weight is the sum of the tip weights and that indicated on the beam.

Automatic indicating scales eliminate the manual operation of placing tip weights and positioning a poise on a graduated beam. In spring-type automatic indicating scales, the steelyard rod is connected to a spring whose deflection is indicated by a pointer against a graduated scale. A dashpot is required to dampen the oscillations of this type of scale.

The primary advantages of mechanical scales are that they are relatively inexpensive in smaller capacities; they do not need electrical power, enhancing their portability; and they do not need to be protected from power line transients or lightning strikes.

Their disadvantages include diminished performance owing to normal wear, damage from corrosives, or buildup of dirt and debris; excessive vibration that may cause damage and render them difficult to read; the fact that for larger sizes the installation is relatively complicated and expensive; their slow speed of operation; and the fact that they do not lend themselves to automation. In order to take advantage of the data collection and automation possibilities afforded by electronics, many older scales have been converted to operate with electronic indicating elements through insertion of a strain-gauge-type load cell in the steelyard rod.

Hydraulic. A hydraulic scale measures weight by supporting the load receiver on the piston of a hydraulic load cell (Fig. 4) and measuring the resulting hydraulic pressure with a Bourdon tube or similar device. Many weighing operations require that the load be supported at three or more points, requiring the use of three or more hydraulic load cells. The individual hydraulic pressures developed by these cells must be converted back to forces for summing and measurement; this is often accomplished using strain-gauge load cells. The hydraulic equipment performs the same function as a mechanical lever system in reducing the load to an easily measured force.

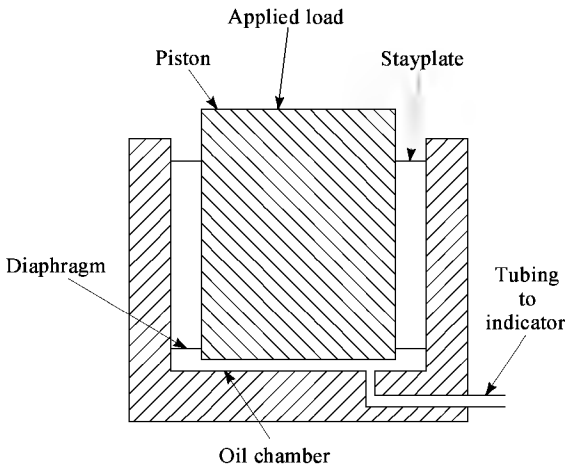


Fig. 4. Hydraulic load cell.

The advantages of a hydraulic scale are immunity from electrical transients and lightning damage, that vibration can be controlled by damping, and that downward travel of the platform or weighing container is relatively small. The disadvantages include the fact that additional equipment, either mechanical or electrical, is required to sum the individual cell outputs in multiple-cell applications, or for control or data acquisition purposes; the fact that the speed of operation is dependent on temperature, ie, as a result of variations in the viscosity of the hydraulic fluid; and the fact that there is a danger of contamination from leaking hydraulic fluid.

Strain-Gauge Load Cells. The majority of industrial scales today use strain-gauge load cells as the weighing element. The strain-gauge load cell is a device which, when a force is applied to it, gives an electrical output proportional to the applied load.

Figure 5 shows a typical metallic foil strain gauge. It consists of an etched grid of very thin foil attached to a thin insulating backing material. The

gauges are bonded firmly to a surface that prevents buckling. When stretched or compressed along the grid lines, the resistance measured between the ends of the grid (solder pads) increases in the case of the former and decreases in the case of the latter.

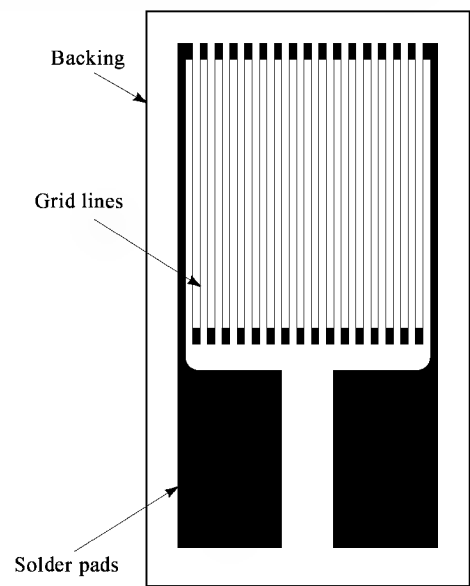


Fig. 5. Foil strain gauge.

The central mechanism of a load cell is the spring element that supports the load, and it is designed to have areas of both tensile and compressive strain suitable for application of strain gauges. Spring elements come in a wide variety of designs, such that choice among them is often dictated by their capacity or the method used in supporting the scale. Figure 6 shows a typical spring element for a moment-insensitive load cell that would be used in a small platform scale. Typically, four gauges are applied in such a way that two are in compression and two in tension (C and T, respectively, in Fig. 6) as the spring element is loaded.

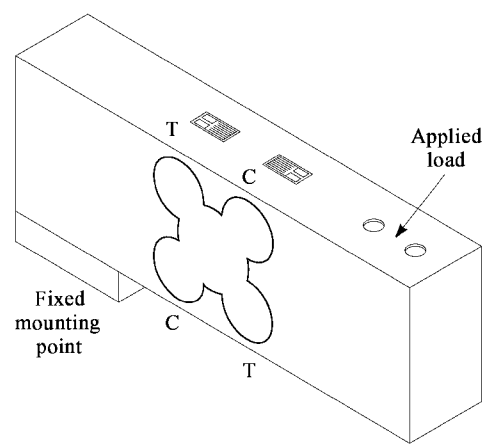


Fig. 6. Load cell spring element. C, gauges in compression; T, gauges in tension.

The four gauges are wired together to form a Wheatstone bridge (3), as shown in Figure 7. An input voltage (typically 10 V dc) is applied as shown, and the resulting output voltage is measured across points A and B. When no load is applied to the cell, all gauge resistances are the same; the bridge is said to be balanced, and the voltage difference between points A and B is zero. As load is applied to the cell, the resistance of the tension gauges increases, whereas that of the compression gauges decreases. The bridge becomes unbalanced, and a voltage difference appears across points A and B. As load continues to increase, the output increases linearly to 20 mV typically at rated capacity.

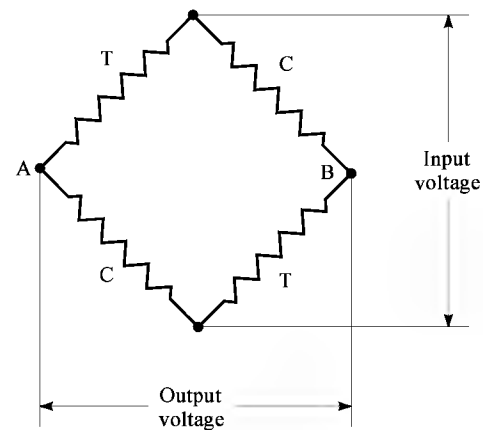


Fig. 7. Wheatstone bridge (see Fig. 6).

Figure 8 shows the basic components of an electronic scale based on strain-gauge load cells. A power supply (not shown, but usually housed in the indicator) provides operating power for the indicator and the input voltage for the load cell. The analogue signal produced by the load cell is sensed by the analogue-to-digital (A/D) converter, which sends digital weight information to the digital computer. The digital computer comprises the microprocessor,

the working storage (RAM), a permanent memory (EEPROM) for storing calibration data, and the program memory (ROM). The computer filters the raw digital signal and processes it into calibrated weight for the display. The weight is also available at the data interface, from which it can be transmitted to any external device. Also, the indicator can be remotely controlled via the data interface; this is an important feature in systems applications.

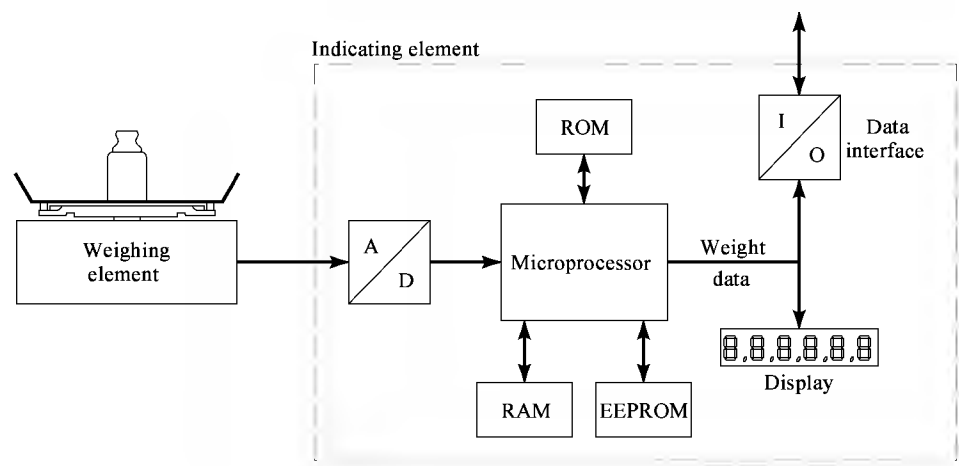


Fig. 8. Basic components of an electronic scale.

- The advantages of scales based on strain gauge load cells are as follows:
- Minimal deflection of the platform or weighing container
 - A large selection of load cells is available in a variety of configurations and materials with capacities from 0.5 kg to 100 t or more
 - Ease of installation; load cells can be used individually or connected in parallel for multiple load-cell applications
 - The weighing and indicating elements may be incorporated into a single housing (as in the case of many bench and retail scales) or the indicating element may be remotely mounted
 - Versatility; the scale components can be purchased individually and used to create custom configurations
 - Availability of a wide range of accessories for data acquisition or process control

- Some disadvantages include the following:
- A very low output voltage from the load cell, commonly 1 μV per displayed division, makes the signal susceptible to noise and degradation (particularly over long distances)
 - Environmental factors such as temperature, humidity, moisture, and Electromagnetic Interference (EMI) or Radio Frequency Interference (RFI) can affect performance if insufficient care is exercised in designing the scale
 - Load-cell performance is limited by properties such as nonlinearity, hysteresis, and creep, which are inherent to the particular design or spring element material.
 - Analogue compensations for temperature effects are required in production
 - Load-cell mountings must be designed in such a way as to minimize the effects of extraneous forces
 - Electrical power is required, along with power-line transient protection

Analogue load cells have remained fundamentally unchanged for over 50 years, but they have been improved greatly in their performance and reliability. Since the mid-1980s, digital load cells have been gaining acceptance in certain applications. A digital load cell has a strain-gauged spring element as described herein, but its output is a robust digital signal. Figure 9 shows a cross section of a typical digital load cell used in truck-scale applications. It consists of a pin-type spring element that is compressed when loaded. The element is surrounded by an enclosure that is welded in place to provide a hermetic seal. The gauges are wired to a printed circuit board which includes an A/D converter and microprocessor; the digital output is through a hermetically sealed connector.

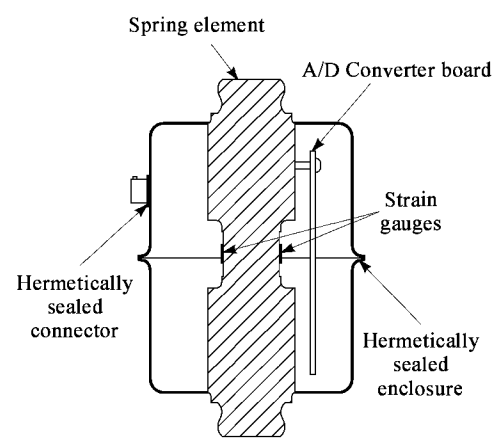


Fig. 9. Digital load cell.

- The advantages of digital load cells compared to analogue are:
- The signal is less susceptible to noise and degradation; this factor is particularly important where long cable runs are necessary
 - Because a microprocessor is built into each cell, compensations for nonlinearity, hysteresis, creep, and temperature fluctuations can be performed digitally to achieve greater consistency and accuracy; also, temperature effects on the A/D are automatically compensated for
 - Outputs can be matched digitally to eliminate laborious corner trimming when installing a scale or when replacing a load cell
 - The indicating element can address each cell independently to display individual load cell weights, or for troubleshooting individual cells.

Electromagnetic Force Compensation. Precision balances traditionally were of the form illustrated in Figure 1. These were very finely made, with agate pivots and bearings and were often housed in an enclosure of glass and highly polished hardwood. Today, the principle of electromagnetic force compensation (EMFC), also referred to as magnetic force restoration (MFR), is used where extremely high accuracy is required. The

fundamental principle of an EMFC balance is that the load added to the pan of the balance is maintained at the same vertical height by varying the current supplied to an electrodynamic converter that supports the pan; the current required is proportional to the load in the pan.

Figure 10 shows a cross section of the weighing cell of a typical EMFC balance. The pan (item 1) is rigidly attached to the hanger (item 2), which is held rigid in a horizontal plane by the parallel guides (item 3). The parallel guides attach to the hanger and fixed chassis (item 4) through the flexures (item 5). This parallelogram arrangement allows the pan and hanger to move freely in the vertical direction, although restraining them completely in the horizontal plane. The hanger is supported in the vertical direction by the lever (item 6); they are attached through the flexible link (item 7). The lever pivots freely at the pivot point (item 8) and has a position indicator (item 9) attached at its free end. The position sensor (item 10) senses the position of the lever, and controls the electrodynamic converter (a device which can convert electrical current into a mechanical force). The electrodynamic converter consists of a permanent magnet (item 11) which produces a uniform magnetic field in the air gap (item 12) between its poles and the coil (item 13) that is positioned in the air gap and attached to the lever. Current flowing in the coil produces a downward force on the lever (4); the position sensor controls the current to maintain the lever at its null position. When a load is added to the pan, the lever deflects. This is sensed by the position sensor, which adjusts the current to the coil such that the lever returns to the null position once again. This current is proportional to the load and is the means by which the mass is determined.

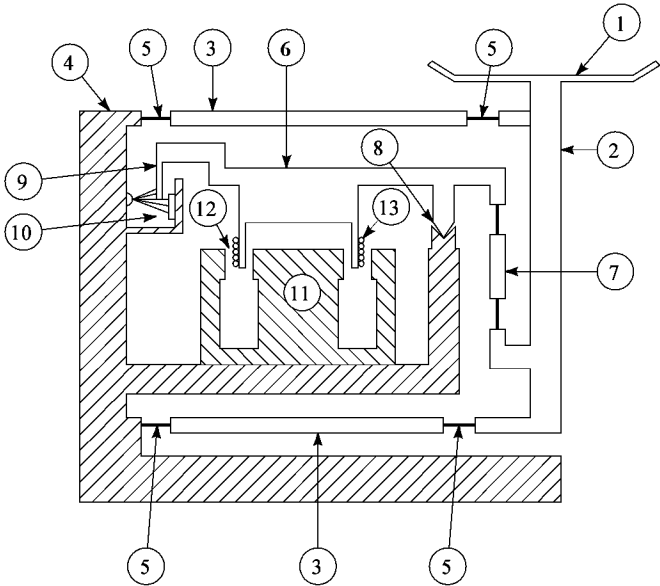


Fig. 10. Cross section of an EMFC weighing cell. Item 1, pan; 2, hanger; 3, parallel guide; 4, chassis; 5, flexure; 6, lever; 7, flexible link; 8, pivot point; 9, position indicator; 10, position sensor; 11, permanent magnet; 12, permanent magnet air gap; 13, compensation coil. See text.

The electronics for a balance based on EMFC technology is very similar to that shown in Figure 8 for a strain-gauge-based scale, with the exception that the weighing element consists of the EMFC cell described herein, together with the circuitry to control the current flow to the compensation coil. Also, the signal sent to the A/D converter is not a voltage, but rather only the current necessary to return the lever to its null position. For balances, the weighing and indicating elements are usually integrated into the same housing.

Because of the extreme accuracy expected of many of these products, some include internal test weights which can be used to recalibrate regularly and to adjust for nonlinearity. Some balances monitor changing conditions and initiate the recalibration procedure as needed.

The advantages of this technology are as follows:

- Extreme accuracy is achievable; balances based on EMFC technology can have up to 50 million displayed divisions, eg, a typical semimicrobalance might have 20.5×10^6 displayed divisions)
- Extremely good repeatability is achievable (as required for mass comparators)
- Faster operation and greater flexibility are achieved, compared to mechanical balances

The disadvantages are as follows:

- Cost, particularly where the highest resolution is required
- These products are very sensitive and are best used in a controlled environment
- Electric power, along with suitable power line transient protection, is required
- Capacity range is limited, although this can be extended through the use of lever systems

Nuclear. Mass can be determined directly by measuring changes in the absorption, reflection, or transmission of alpha- or beta-rays, which changes in proportion to the amount of material present. This method is primarily used to determine the mass of bulk material moving on a conveyor. The advantages include the following:

- Simple installation
- No contact with the material, no moving parts to wear out or corrode
- Unaffected by changes in the tension or stiffness of the conveyor belt
- Direct readout and adaptability to modern controls

Some of the disadvantages are as follows:

- Low accuracy, unsuitability for commercial transactions
- Variations in material geometry and density affect accuracy
- Requires shielding and special procedures to avoid exposure of personnel to radiation
- Frequent recalibration is required to compensate for source decay over time
- Electrical power and power line transient protection required

Commercial Regulatory Aspects and Performance Characteristics

Scale Performance. Figure 11 shows the relationship between the applied load and the scale reading. The ideal scale would have a perfectly linear relationship, as represented by the straight line, and the results from increasing and decreasing load tests would fall on this line. The two broken lines represent the typical calibration curve (greatly exaggerated); the arrows indicate increasing and decreasing load directions.

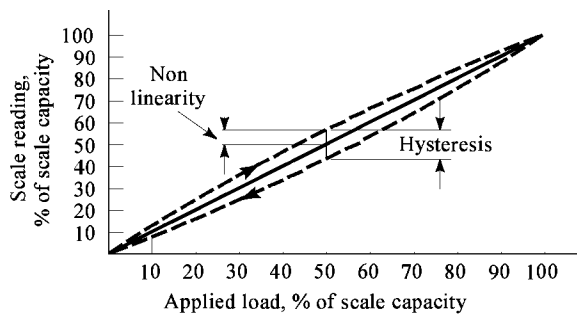


Fig. 11. Scale calibration curve (---); —, increasing load direction; —, decreasing load direction.

- Definitions of some terms commonly used in describing scale performance follow (6,7):
- Nonlinearity*: the maximum deviation from the straight line as load is applied, usually expressed as a percentage of scale capacity (see Fig. 11)
 - Hysteresis*: the maximum difference between the increasing and decreasing load curves at a given load, expressed as a percentage of scale capacity (see Fig. 11)
 - Nonrepeatability*: the maximum difference in scale readings when the same test load is applied repeatedly under the same conditions, expressed as a percentage of scale capacity
 - Creep*: the change in scale reading when a load (usually scale capacity) is applied for a period of time, and all other variables are held constant; it is expressed as a percentage of applied load in a specified time period
 - Value of the scale division, d*: the difference between two consecutive weight readings on the scale's display, expressed in units of mass, eg, if a scale display increments upwards in steps of 0.02 g, $d = 0.02\text{ g}$
 - Number of scale divisions*: the scale capacity divided by d , eg, a scale of 5-kg capacity with a 1-g scale division has 5,000 scale divisions
 - Shift error*: the range of results obtained when a test weight is moved around the scale platform in a specified manner (5), expressed as a percentage of scale capacity
 - Temperature effect on zero*: the change in the scale's no-load reading with changes in ambient temperature, expressed as a percentage of scale capacity per °C, or the number of scale divisions per 5°C
 - Temperature effect on span*: the change in a scale's calibration with changes in ambient temperature, expressed as a percentage of scale capacity per °C
 - Accuracy*: a measure of the extent to which a scale can provide results that conform to the true value

U.S. Regulations. Weighing equipment used in commercial transactions is subject to regulation and inspection by state or local weights and measures agencies. It must comply with certain design and performance requirements, and must have a certificate of conformance issued prior to installation. *Handbook 44* issued by the National Institute of Standards and Technology is the basis for weights and measures enforcement activities in the United States (5). It covers scales used in buying or selling goods of any kind, and scales for determining the cost of services, eg, shipping and laundry charges. Although *Handbook 44* does not apply to scales used only to control processes or to measure yield or inventory, the majority of scales used comply with its requirements. Thus, *Handbook 44* provides a standardized and comprehensive basis for comparison of scales for industrial and retail applications.

Handbook 44 defines five accuracy classes for scales in terms of the value of the scale division and the number of divisions. Class I applies to precision laboratory weighing. Class II applies to laboratory weighing (precious metals, gems, and grain test scales). Class III applies to the majority of industrial and retail scales, and to all scales not specified in the other categories. Class III L applies to vehicle, livestock, railway, crane, and hopper scales. Class IIII applies to portable scales used for highway weight enforcement.

Table 1 is condensed from *Handbook 44*. It lists the number of divisions allowed for each class, eg, a Class III scale must have between 100 and 1,200 divisions. Also, for each class it lists the acceptance tolerances applicable to test load ranges expressed in divisions (d); for example, for test loads from 0 to 5,000 d , a Class II scale has an acceptance tolerance of $\pm 0.5\text{ d}$. The least ambiguous way to specify the accuracy for an industrial or retail scale is to specify an accuracy class and the number of divisions, eg, Class III, 5,000 divisions. It must be noted that this is not the same as 1 part in 5,000, which is another method commonly used to specify accuracy; eg, a Class III 5,000 d scale is allowed a tolerance which varies from $\pm 0.5\text{ d}$ at zero to $\pm 2.5\text{ d}$ at 5,000 divisions. Calibration curves are typically plotted as in Figure 12, which shows a typical 5,000-division Class III scale. The error tunnel (stepped lines, top and bottom) is defined by the acceptance tolerances listed in Table 1. The three calibration curves belong to the same scale tested at three different temperatures. Performance must remain within the error tunnel under the combined effect of nonlinearity, hysteresis, and temperature effect on span. Other specifications, including those for temperature effect on zero, nonrepeatability, shift error, and creep may be found in *Handbook 44* (5). The acceptance tolerances in Table 1 apply to new or reconditioned equipment tested within 30 days of being put into service. After that, maintenance tolerances apply; they are twice the values listed in Table 1.

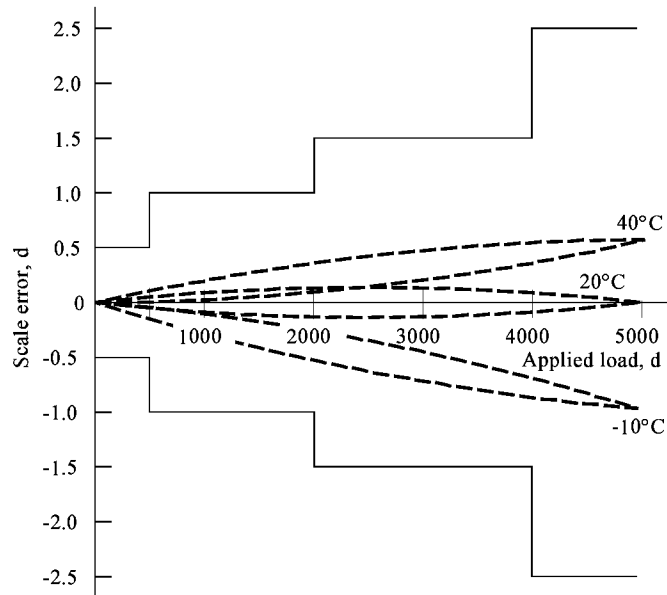


Fig. 12. Calibration curves for a *Handbook 44* Class III 5,000 d scale. See text.

Table 1. Summary of Tolerances and Number of Divisions Applicable to Accuracy Classes^a

Parameter	All values in this table are in scale divisions				
	Class I	Class II	Class III	Class III L	Class IIII
allowed number of divisions	50,000 ±	100–100,000	100–10,000	2,000–10,000	100–1,200
acceptance tolerances					
±0.5 d	0–50,000	0–5,000	0–500	0–500	0–50
±1 d	50,001–200,000	5,001–20,000	501–2,000	501–1,000	51–200
±1.5 d	200,000 ±	20,001–100,000	2,001–4,000		201–400
±2.5 d			4,001–10,000		401–1,200
add 0.5 d 500 d ^b				1,001–10,000	

^a Ref. 5.
^b Add 0.5 d tolerance for each additional 500 d or fraction thereof.

Regulation Outside the United States. Each country establishes its own weights and measures requirements. The majority of these are based on the recommendations of the Organisation Internationale de M  trologie L  gale (OIML), in Paris. R76-1 is the OIML equivalent of *Handbook 44*; it uses accuracy classes and an acceptance tolerance structure similar in many ways to Table 1 (8).

Factors Affecting Weighing Accuracy

Variations in the Force Due to Gravity. The mass of an object is the quantity of matter in the object. It is a fundamental quantity that is fixed, and does not change with time, temperature, location, etc. The standard for mass is a platinum–iridium cylinder, called the International Kilogram, maintained at the International Bureau of Weights and Measures, in S  vres, France. The mass of this cylinder is 1 kg by definition (9). All national mass standards are traceable to this artifact standard.

It is easiest to appreciate what is meant by mass indirectly, by observing the influence of forces on objects, eg, by picking up an object and sensing the effect of the earth's gravitational force acting on it, and hence "feeling" its weight.

Newton's law of gravitation states that if two particles are a distance *r* apart, the mutual attraction force between them can be expressed as follows (10):

$$F = \frac{6.673 \times (10^{-11}) m_0 m}{r^2}$$

(1)

where *F* is the mutual attraction force in N, *m*₀ and *m* are the masses of the particles in kg, and *r* is the distance between their centers in meters. This equation can be used to calculate the attractive force between the earth and a mass, *m*, resting on its surface (mass of the earth is 5.976 × 10²⁴ kg and its mean radius is 6,371 × 10³ meters) as follows:

$$F = \frac{6.673 \times (10^{-11}) 5.976 \times (10^{24}) (m)}{(6,371 \times 10^3)^2} = 9.82(m)$$

(2)

A mass, *m*, resting on the surface of the earth is attracted to it with a force of 9.82(*m*) N. Put another way, the force per kg experienced by a mass resting on the earth is 9.82 N/kg. The proportionality factor 9.82 is referred to as gravity, or *g*; it is sometimes referred to as acceleration due to gravity with the equivalent units of m/s² (11).

The factor *g* was calculated using the earth's mean radius; *g* actually varies with changes in location and altitude. The earth is not spherical; it bulges at the equator; ie, an object at either of the poles is actually nearer to the center of the earth. Also, the earth is rotating on its axis, and an object at the equator is subject to a small centrifugal force that is not present at the poles. For both of these reasons, *g* increases with increasing latitude from the equator to the poles. In addition, *g* decreases with increasing altitude above the earth's surface. All of these factors are taken into account in Helmert's equation for *g*:

$$g = 9.80616 - 0.025928 \cos 2\phi + 0.000069 \cos^2 2\phi - 3.086 \times 10^{-6} h$$

(3)

where ϕ is the latitude in degrees and *h* is the altitude in meters. Table 2 gives the value of *g* at various locations.

Table 2. Values for Gravity, *g*, at Various Locations

Location	Gravity, <i>g</i> N/Kg	Location	Gravity, <i>g</i> N/kg
<i>United States</i>		<i>Other countries and areas</i>	
St. Michael, Alaska	9.82192	Arctic Red River, Canada	9.82434
Quiet Harbor, Alaska	9.81624	Vancouver, Canada	9.80949
Los Angeles	9.79595	Canal Zone, Panama	9.78243
San Francisco	9.79965	Greenwich, U.K.	9.81188
Denver	9.79609	K��nigsberg, Germany	9.81477
Hartford, Conn.	9.80336	Rome, Italy	9.80367
Washington, D.C.	9.80095	Monrovia, Liberia	9.78165
Key West	9.78970	Bergen, Norway	9.81922
Miami	9.79053	Stockholm, Sweden	9.81843
Honolulu	9.78591	Bah��, Brazil	9.78331
St. Louis	9.80001	Kingston, Jamaica	9.78591
New York	9.80267	North or South Pole, sea level	9.83217
El Paso	9.79124	equator, sea level	9.78039
Seattle	9.80733	equator, 500 m above sea level	9.76496

Using an equal-arm balance (see Fig. 1), the unknown mass of an object can be determined by placing it in one pan, and adding test weights to the other until the beam balances. The result will be the same regardless of location because the object and the test weights are subject to the same value of *g*. Any scales that measure an unknown mass by comparing it with a known mass (with or without a lever system), will be unaffected by variations in *g*.

However, this is not a very convenient method of weighing, and the majority of scales in use today determine the mass of an object indirectly by measuring the gravitational force acting on it. If a scale is calibrated at one location, and is used at another, the reading would change by a factor of g/g_c , where g and g_c are gravity at the point of use and calibration, respectively. If a mass weighed 5,000 kg on a force-measuring scale in Miami, Florida, it would weigh 5016 kg (5,000 kg $\times$ 9.82192 / 9.79053) if weighed on the same scale in St. Michael, Alaska, a change of 0.3%. In the extreme case, a variation of 0.7% could be expected. Force-measuring scales should be recalibrated when relocated.

Some scales have a software feature whereby a certain geographical region is identified on initial configuration, and the software then makes the appropriate correction. Some high performance balances have built-in test weights for initial (and periodic) calibration.

Scales are sensitive to force applied in one direction only, eg, a scale with a horizontal platform is sensitive to forces applied perpendicular to the platform. Scales should be leveled before calibration and whenever they are moved; portable scales generally have a bubble level to facilitate leveling.

Slight changes in altitude can significantly affect precision balances of the force-measuring type (note that Helmert's equation for g accounts for the effects of altitude). For example, if an analytical balance is relocated from the basement to the 12th floor of the same building (a change of 48 m in altitude), and is used to weigh a mass of 250 g, the reading will be lighter by 3.8 mg on the 12th floor. Again, recalibration is required.

Moisture Content. In making very fine measurements, it might be necessary to protect a sample from the possibility of losing or absorbing moisture during the weighing operation. This is particularly true if the weighing chamber is at a different temperature or humidity compared to the sample's storage location. In weighing highly volatile liquids, evaporation during the weighing process may cause a steady decline in the weight reading. In either case, the sample should be weighed in the smallest practical covered container.

Materials subject to variable moisture content are frequently treated on a dry weight basis, or on some other standard moisture content basis. In such cases, the moisture content of the material being weighed must be determined, and the actual weight converted to the desired basis. Moisture analyzers are available that incorporate a balance and heating element in one unit. They monitor a sample's weight while the temperature is elevated to expel moisture. Typically, such units operate automatically and display the moisture content directly as a percentage of the original weight.

Temperature Effects. The output of a strain-gauge load cell changes with temperature because of, for example, changes in gauge resistance and because the gauge and spring element have different coefficients of thermal expansion. Also, with EMFC balances the strength of the permanent magnet changes with temperature, changing the amount of current necessary to balance the load. Good design and careful manufacturing can compensate for most of these effects, but they will not eliminate them completely.

When power is applied to an electronic scale, it warms up and is somewhat unstable until thermal equilibrium is reestablished. The manufacturer's recommendations with regard to warm-up period should be followed.

Strain-gauge load cells are sensitive to temperature gradients induced by, for example, radiant heat from the sun or resulting from high temperature wash down. Load cells should be shielded from such effects or given time to stabilize before use.

Scales will always perform better in a temperature-controlled environment; this is particularly important for high precision balances. Those designed to automatically recalibrate as conditions change will perform better in less-than-ideal conditions.

Where there is a temperature difference between the object to be weighed and the surrounding air, air currents will be induced close to the object's surface (12). These can be significant if extreme accuracy is required. Objects should be allowed to reach thermal equilibrium in the laboratory before weighing. Just as important, the balance should be designed to minimize the temperature rise inside the weighing chamber. In extreme cases, the object should be placed inside the chamber until it reaches thermal equilibrium before weighing. Needless to say, drafts must be avoided.

Buoyant Effect of Air. Weighing operations performed *in vacuo* are not affected by buoyancy forces. An object in air, however, is subject to a buoyancy force that is equal and opposite to the gravitational force on the mass of air the object displaces (10). If the equal arm balance of Figure 1 is in balance with a test weight of mass, m' , in one pan, and material of mass, m , in the other, $m = m'$ if they have the same density. If the densities are different, then the buoyancy forces acting on each pan affect the result. Taking moments about the center pivot point gives

$$g(m - \rho_{\text{air}} V_m) = (m' - \rho_{\text{air}} V_w) g \tag{4}$$

where ρ_{air} is the density of air (1.2 kg / m³ on average), V_m and V_w are the volumes of the material and test weight, respectively, in m³. By expressing V_m and V_w in terms of mass and density, this equation can be expressed as

$$m - m'(1 - \rho_{\text{air}} / \rho_w) = (1 - \rho_{\text{air}} / \rho_m) m \tag{5}$$

where ρ_m and ρ_w are the densities of the material and test weight, respectively, in kg / m³. This expression can be used to calculate the actual mass, m , from the apparent mass, m' , indicated by a scale (true for all scales except nuclear). From equation 5, it can be seen that if $\rho_m = \rho_w$, then $m = m'$, as would be expected.

If a 1-kg stainless weight ($m' = 1,000\text{g}$, $\rho_w = 8,000\text{ kg / m}^3$) is added to one pan of the balance in Figure 1, and material with a density of 1,000 kg / m³ is added to the other until equilibrium is reached, the amount of the material needed is 1001.05 g, using equation 5. Thus, it takes 1001.05 g of this material to counterbalance 1,000 g of stainless steel, because of the buoyancy effects on the dissimilar volumes.

Buoyancy has no effect in weighing material of the same density as the test weight (used for direct comparison, or used to calibrate the balance), or in proportioning materials of approximately the same density. In general, the recipes used in proportioning are derived empirically, and account for the effects of buoyancy. Buoyancy effects can be significant when the absolute mass must be known accurately, eg, in a laboratory environment, and corrections can be made using equation 5. In industrial and commercial weighing, the effect of buoyancy is generally not considered. Note that variations in air density, eg, due to variations in temperature, can also have a slight effect on weighing repeatability.

Electrostatic and Magnetic Effects. These two effects are generally small but may be significant in laboratory weighing.

If an electrostatically charged object is weighed close to a charged or conducting surface not being weighed, then weighing accuracy is likely to be affected by the forces which result. Similar problems arise as a result of magnetic effects. Some materials can be permanently magnetized, and others temporarily exhibit magnetic behavior when placed in a magnetic field. When weighed, these materials can seem heavier or lighter if they are in close proximity to a magnetically permeable material, or if they are surrounded by a magnetic field, eg, the earth's magnetic field, or one produced by external equipment.

To avoid electrostatic problems, the air in the laboratory should be maintained at 40% relative humidity or higher, or the air should be ionized. Shielding the sample is also effective against electrostatic and magnetic influences. If the item is electrostatically charged or in a magnetized state, it should be weighed in a metal container. The container can be any conductive metal for electrostatic problems, but it should be magnetically permeable (such as soft iron) for magnetized items. Also, an item that is strongly magnetic should be placed on a nonpermeable spacer such as plastic to lift it away from the scale.

A slightly different arrangement is called for if the balance becomes electrostatically charged, which may happen to the glass panels of the draft shield in low humidity environments, for example, or if it is subjected to an external magnetic field. In this case a shield of the materials described above should be inserted between the item on the balance and the external influence. The shield, eg, a bell-shaped container, must not be supported by the pan of the balance. To facilitate shielding, some balances have an opening which allows a weighing pan to be suspended at any distance below it.

Types of Scales

Scales are available in a variety of designs and configurations to facilitate different weighing operations (7). The two principal categories are industrial and retail scales, and precision scales and balances.

Industrial and Retail Scales. Scales using strain-gauge load cells predominate in this market segment, although mechanical and hydraulic

scales are also used to some extent. Many are used in commercial applications, and are typically of accuracy class III or III L with up to 10,000 divisions. The following are descriptions of some industrial and retail scales.

Bench scales are small and light enough to be placed on a bench or table. They generally have a platform on which small quantities can be placed by hand. Alternatively, some have a roller conveyor platform on which the item is moved across the scale. Capacities range from 5 kg to 250 kg.

Retail scales in various configurations are used in stores and supermarkets to determine price. They range from simple weigh-only scales to ones having touch screens and enough memory to store and print data on thousands of products, including the data necessary to print nutritional labels. They range in capacities up to 25 kg.

Portable scales generally have a platform near the floor with the indicator mounted on a column for ease of operation. These scales may be on wheels, and many are battery-powered. Capacities are generally less than 600 kg. Bench scales can be placed on a wheeled table for similar use.

Counting scales are a form of bench scale that can display not only weight, but also the number of parts on the weighing platform. In operation, a small number of parts is put on the scale, which then automatically establishes the average piece weight; if a large quantity is placed on the scale, the scale divides the total weight by the average piece weight and displays the number of parts.

Floor scales are platform scales which are permanently installed in a pit, flush with the floor, or which sit on the floor. These scales typically range in capacity from 500 kg to 30 t, and in size up to that required to accommodate a fork-lift truck.

Truck scales weigh highway vehicles; they may be installed in a pit or above-ground with inclined approach ramps. A capacity of 100-t is typical. For monitoring axle weights, the scale deck can be broken into three scales, which are monitored independently; the individual axle weights or total vehicle weight can be displayed by the indicator.

Railroad-track scales are installed in a pit and support the appropriate length of rail to weigh railroad cars. For static weighing of cars, the rails are generally long enough to support the entire car; capacities range up to 350-t. Some railroad weighing is done with the cars coupled and in motion (CIM). In this case the rails of the scale may be long enough to weigh the entire car or just each truck. Some scales have a flat deck so that highway vehicles can be weighed in addition to railroad cars.

Monorail scales are constructed by removing a section of monorail track and replacing it with a similar section that is mounted on a scale. Loads transported by crane or on roller hangers can be weighed quickly and easily. Capacities range up to 5 t.

Crane scales are designed to be inserted between the hook of a crane and the load being lifted. The upper end of the scale has an eye which attaches to the crane's hook, whereas the lower end has a hook that picks up the load. A tension-type load cell connects the eye and hook. The display may be either integral with the load cell or remotely mounted. Capacities range up to 50 t.

Tank and *hopper scales* weigh liquids or bulk materials. The materials may be stored in the scale, or may be weighed while being transported for sale or for use in a process. The scale is usually part of a much larger material handling system. Capacities are mostly under 20 t, but can be as large as 200 t or more.

Drum-filling scales position a container on the weighing platform and, using a suitable lance, fill the container to the required weight. The handling and filling of the containers can be either manual or fully automated. *Bag-filling scales* are of a similar type; the material may be weighed either in the bag or in a hopper scale before discharge into the bag.

Belt-conveyor scales determine the amount of material being conveyed on a belt. A section of belt is weighed by placing the belt support rollers on a scale; the belt speed is also measured. Weight and speed data are supplied to a controller which integrates them to arrive at a material flow rate, often stated in tons per hour. The controller may display a flow rate, shut the conveyor down when a predetermined amount of material has passed, or it may be used to maintain a specified flow rate. Accuracy is limited because of the number of detrimental influences involved, eg, variable belt tension.

In-motion checkweighers weigh discrete items as they move along a production line on a belt or chain. The checkweigher may automatically eject items of unacceptable weight, or may divert them into several streams corresponding to predetermined weight categories. The weight information may be analyzed statistically and used to adjust the filling equipment automatically. In certain applications, the checkweigher may regulate the spacing of packages before weighing. Systems are available to handle various package sizes at speeds of up to 900 packages per minute.

Precision Scales and Balances. These products today rely predominantly on electromagnetic force compensation (EMFC) technology. Most are not used in commercial transactions; however, many approved models are available. Some of the precision scales fall into accuracy class III, but Class I and Class II are more typical for the balances. *Handbook 44* is of limited value in describing the accuracy characteristics of these products because, for example, 1 mg is the smallest division size allowed. For this reason, only the number of displayed divisions is discussed herein.

Types of precision scales and balances include the following:

Precision bench scales are designed for use in industrial settings; they typically have capacities below 50 kg and 30,000 or more displayed divisions. Typical applications are counting very small parts or weighing hand-add materials for proportioning operations.

Precision floor scales have platform sizes of 1.5 m square or larger, and typically use a lever system to transmit the load to a single EMFC load cell. Capacities range up to 6 t with a minimum of 30,000 displayed divisions.

Precision balances are similar to precision bench scales but are typically used in laboratory settings with up to 500,000 displayed divisions.

Analytical balances are higher accuracy laboratory balances with capacities up to 400 g, and up to 20 million displayed divisions.

Microbalances typically have capacities from 2 g to 5 g, with 20 to 50 million displayed divisions.

Mass comparators are balances of particularly high accuracy and are designed, as the name implies, for comparison of masses within a very narrow weighing range. A typical application is the calibration of test weights by comparison to known standards. Repeatability of weighments is the critical factor; a typical comparator might have a capacity of 1,000 g, a displayed division size of 1 µg and a repeatability of 2 µg (one part in 500 million). Capacities range from 5 g to 1.5 t.

Weighing Methodologies

Scales can be portable so that they can be moved about and temporarily located where the weighing operation is to take place; they may be light enough to be carried by hand, or they may be on wheels or be transported using a lift truck. Even some truck scales are considered portable because, for example, they are designed to be transported from one highway construction project to another. Some scales are attached or built into portable material-handling devices such as cranes and lift trucks, so that the weighing operation can take place while material is being transported.

Some scales are fixed in position and the material to be weighed is brought to the scale. Higher capacity floor scales, truck scales, and railroad-track scales are typically fixed in position; these scales should be located where they are convenient for the weighing operation, but are not subjected to unnecessary abuse from through traffic. Scales are often built into fixed material handling or processing equipment so that material can be weighed as it is being moved, eg, as in the cases of conveyor, tank, and hopper scales. Scales can be built into tanks or silos used for storage, either for inventory control or for controlling the discharge of these materials.

The object of a weighing operation may be to determine the gross weight (the weight of the material plus the tare weight of its container) or, more commonly, the net weight (the weight of material exclusive of its container). Gross weight is used to determine shipping charges, for example, while net weight is used to determine selling price.

There are various reasons for weighing materials and various methods employed; the most common ones are discussed below. It is assumed for the sake of discussion that net weight is the requirement, since it is the more common.

Weighing Predefined Quantities. In this form of weighing, the amount of material has been defined prior to weighing, eg, by filling a truck to a certain level, and the weighing operation is carried out simply to determine what the weight is. An object may be placed on a scale and its net weight will be displayed directly (assuming the object is not packaged). Bulk materials may be handled in anything from small containers to railroad cars; the weighing operation consists of measuring the gross weight and subtracting the container's tare weight to arrive at the material's net weight. A truck in-out operation provides a good illustration of this form of weighing; when a truck arrives at a terminal, it pulls onto the scale and its weight (gross or tare,

depending on whether it is delivering or picking up a load) is stored in memory by the scale. The truck is then filled or emptied and returns to the scale before exiting the terminal; the scale now has the truck's gross and tare weights and can automatically display the net. Typically the scale prints a ticket listing gross, tare, and net weights, along with the time and date. The information may also be transmitted to a computer for accounting or inventory control.

Tank and hopper scales can weigh in one draft or in multiple drafts. A hopper scale used for multiple-draft weighing of material in transit is commonly called a bulk weigher (Fig. 13). An accumulating or surge hopper with a feed gate is installed above the scale hopper; a surge hopper is also installed below the scale hopper. In operation, material flows into the scale until the feed gate is closed by the controller; the actual load on the scale is recorded and the scale hopper's discharge gate is opened temporarily to empty the scale. Any material left in the scale is weighed and subtracted from the stored weight to arrive at the net weight for that draft. This operation is repeated until the quantity to be weighed has passed through the scale. This system can be used to weigh a definite amount, such as that contained in a railroad car or a ship, with an error of $<0.1\%$ of net weight. If the input is from a car or ship and the output is to storage, it functions as a receiving scale; if the input is from storage and the output to a car or ship, it functions as a shipping scale. As a receiving scale, it weighs a predefined quantity; as a shipping scale, it may be used to weigh out a particular quantity to fill a car or ship optimally.

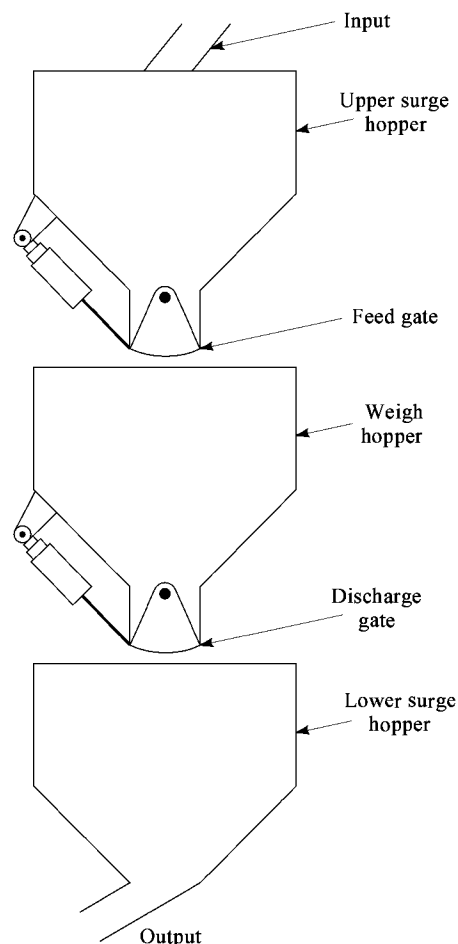


Fig. 13. Bulk weigher.

Some of the various methods of weighing the material in a railroad car, arranged approximately in increasing order of accuracy, are:

- Weighing the loaded car on a coupled-in-motion track scale. An entire train, typically, is hauled at slow speed over the scale. The net weight is determined by subtracting the tare weight stenciled on the car
- Weighing the material on a belt-conveyor scale as it is put into, or removed from, the car
- Weighing the gross weight of the car on a static railroad track scale and subtracting the actual tare weight of the car, determined by weighing the empty car before loading or after emptying, as appropriate
- Using a bulk weigher as described above to measure the material either being placed in, or removed from, the car
- Weighing the complete load delivered to, or received from, the car as a single draft in a hopper scale.

Controlling Weight. Filling containers to predetermined weights is the most common example of controlling weight. One method is the net-weighing process, in which material is fed into a hopper or tank scale until the desired weight is reached; it is then discharged into the container. This method is often used in bag filling. In the gross-weighing method, the container is placed directly on the scale and filled to the required weight. Drum fillers are an example of the use of this method. Either filling operation may be controlled manually by an operator observing the scale display; more typically, the filling is completely automated for better consistency and greater speed.

Continuous Weighing and Controlling. Scales can be used to determine the weight per unit length of material in sheet or strip form as it is being manufactured or transported. This weight information can be used to control moisture content, addition of sizing or other materials, or to control an extrusion process. The material can pass over a roll that is weighed, or over a belt-conveyor scale. Direct mass measurement with a nuclear scale can also be used for this application.

Process industries frequently need to weigh and control the flow rate of bulk material for optimum performance of such devices as grinders or pulverizers, or for controlling additives, eg, to water supplies. A scale can be installed in a belt conveyor, or a short belt feeder can be mounted on a platform scale. Either can be equipped with controls to maintain the feed rate within limits by controlling the operation of the device feeding the material to the conveyor. Direct mass measurement with a nuclear scale can also be used to measure and control such a continuous stream of material.

Two weigh hoppers can be arranged in parallel to provide accurate quantities of material without interruption (Fig. 14). In operation, weigh hopper A discharges while weigh hopper B is being filled to the appropriate weight. When hopper A empties completely, hopper B begins discharging; hopper A begins filling again, ready to discharge when hopper B is empty. Through correct sequencing, a continuous stream of material can be supplied.

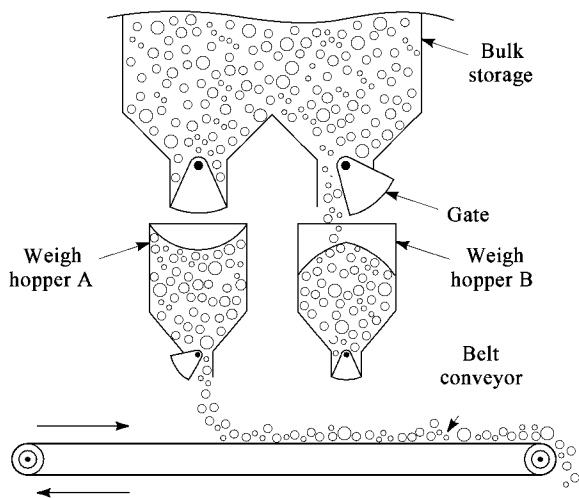


Fig. 14. Parallel weigh hoppers.

The most accurate flow rate control can be achieved by using the loss-in-weight method. The total amount of material required for a downstream process is first added to a tank or hopper scale. As the material is discharged, the loss-in-weight is monitored and used to modulate the discharge valve or gate to achieve the desired flow rate.

Material Proportioning. In proportioning operations, two or more materials are weighed and mixed according to a recipe for use in a chemical process or to make a mixed product, such as animal feed. Proportioning can be performed manually on a bench scale or automatically in tank and hopper scales employed with elaborate material handling systems and controls. Proportioning can be done on a continuous basis or, more typically, in batches, which yield greater accuracy.

Where the manufacturing process is continuous, the equipment described herein on continuous weighing can be used. For low accuracy systems, two or more belt-conveyor scales can be run in parallel, supplying different materials at appropriate feed rates. For higher accuracy, two or more parallel weigh hopper systems (see Fig. 14) can be sized appropriately for each material and run in parallel to provide continuous proportioning. Multiple loss-in-weight systems may also be used. In some processes, the flow of the primary ingredient is measured but not controlled, and the addition of minor ingredients is modulated accordingly for correct proportioning, using a belt-conveyor scale or loss-in-weight feeder.

The three principal methods of batch proportioning are accumulative, sequential, and simultaneous; the best method for a given application depends on the processing equipment available and the accuracy required.

In the case of accumulative proportioning, a single-weigh hopper having adequate capacity for the entire batch is used (Fig. 15). Each ingredient is weighed in turn and accumulated in the hopper scale until completion of the batch. This system has the advantage of lower cost and space requirements; the disadvantage is that the accuracy of minor ingredients suffers if a small amount of these materials is weighed relative to the scale's capacity. Also, in multiple-recipe systems there is the danger of cross contamination if some materials are not used in all recipes.

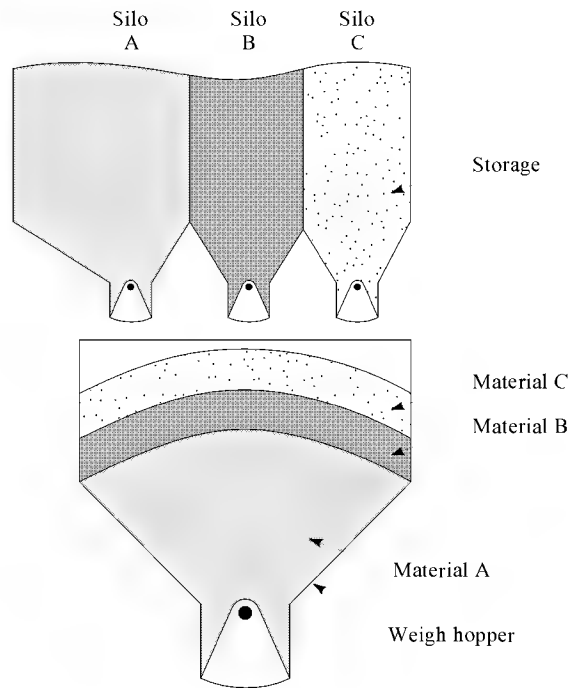


Fig. 15. Accumulative proportioning system.

In sequential proportioning, the physical configuration is the same as that shown in Figure 15, but each material is weighed and discharged before weighing of the next ingredient begins. If all ingredients weigh about the same, higher accuracy can be achieved with this system by matching the weigh hopper capacity with the weight of material to be weighed. This system is relatively slow because of the multiple-discharge cycles.

In the case of simultaneous proportioning, a weigh hopper is provided for each material, as shown in Figure 16. This system has several advantages. It is fast because all materials are weighed and discharged simultaneously. Before discharging, the weight of each ingredient can be compared to the recipe and adjusted if necessary. There is less danger of contamination if some materials are not used in all recipes. The capacity of the weigh hoppers can be matched closely to the weight of each ingredient. By simultaneous discharge onto the belt, a certain amount of premixing can be achieved. The disadvantages are the relatively high cost and the amount of space required. Simultaneous systems can be designed to handle up to 150 tons per hour or more, and are often used to feed several mixers or processes. Combination systems are becoming more common: in this arrangement, an accumulative hopper is included in a simultaneous proportioning system. Individual weigh hoppers are provided for the high volume and critical materials, whereas several of the lower volume or less critical materials are proportioned in the accumulative weigh hopper. Assuming proper sizing of such a system, cost is reduced without compromising speed. Regardless of the system used, critical ingredients such as food additives or dyes are often weighed on an

appropriate balance and added manually.

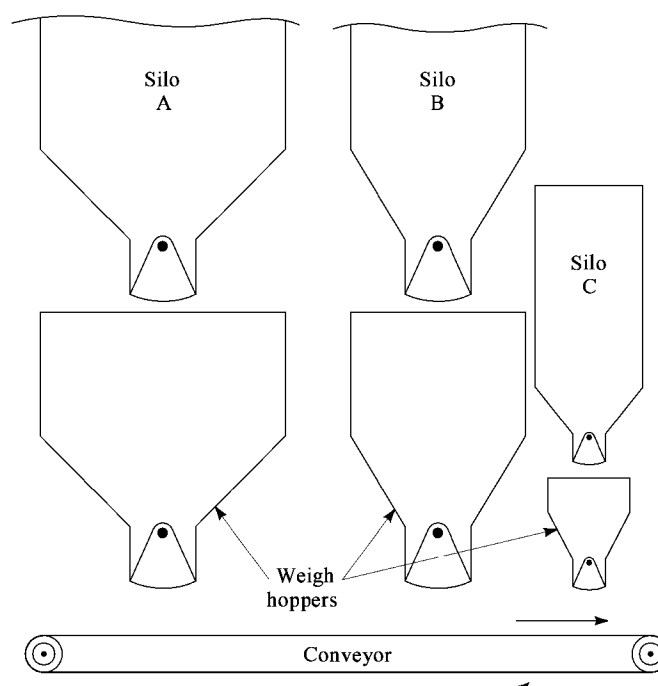


Fig. 16. Simultaneous proportioning system.

Scale Functionality

Mechanical scales indicate the weight applied to the scale through the manipulation of a poise weight on a beam, or automatically using a pointer and dial. Simple control functions can be accomplished by adding position sensors to the indicator or lever system. However, many mechanical scales have been converted to electromechanical by inserting a load cell in the steelyard rod; this allows the scale to take advantage of the wide range of functionality available from electronic indicators. In new installations, electronic scales are typically installed if any form of data collection or control is required.

Electronic scales are microprocessor-based devices that provide a range of features for various applications. At its most basic, an electronic scale provides a digital display of the applied weight; a zero button is usually provided so that slight deviations from the zero indication can be corrected periodically. Many simpler scales also provide tare and gross net buttons. The tare button is used to store an empty container weight in memory for subtraction from subsequent weighments; the gross net button allows the display to toggle between the gross and net weights. Where it is equipped with numeric or alphanumeric keypads and various function keys, an electronic scale can perform a number of functions directly or indirectly related to weighing. The following examples illustrate some possible uses.

Some indicators are designed for truck in-out operations. The indicator can store tare weights for several hundred vehicles, linking them to the trucks' registration numbers. Regardless of the sequence in which the trucks arrive and depart, the scale can provide a printout of gross, tare, net, and time and date for each transaction with just a single weighing of each vehicle. Net weights can also be accumulated in user-specified categories to provide totals and subtotals for billing or inventory control.

Counting scales display the number of parts on the scale as well as the total weight. They establish the average piece weight (APW) of the parts to be counted, which can be stored in memory along with the part number and an associated tare weight. In the case of cycle counting operations, for example, a bin of parts is placed on the scale and the APW and container weight is recalled by keying in the part number; the scale then calculates the net weight of the parts and divides by the APW to arrive at the parts count. This information can be displayed and stored in memory for automatic uploading to the main computer for correction of inventory records. Counting scales can also be used with bar code scanners to eliminate the need for keying part numbers. It is common practice to store all APW and tare weights on a main computer from which it is downloaded to the scale daily, as an indication of what needs to be cycle counted that day.

Over under scales are often used for manually filling containers to a desired weight; they are often found in the meat and poultry industry. In addition to the normal weight display, these scales typically have lights or a bar graph to indicate to the operator when the weight is under, within, or over the desired weight range. Employing a part or product look-up (PLU) number, various information can be stored for several items, eg, the target weight, the allowable tolerance values around this target, and an associated tare weight. This information can be programmed at the scale or downloaded from a central computer. Groups of these scales are often networked together with a central computer which can be used to download setup parameters to the scales for fast production line changeovers. The central computer can also collect weight information from each scale for production monitoring and statistical process control.

Scales can be connected with printers, remote displays, bar code scanners, keyboards, relays, computers, and various controllers such as programmable logic controllers (PLCs). Scales typically have one or more simplex or duplex serial ports that can be current loop, RS-232, RS-422, or RS-485. Frequently ports are programmable for on-demand or continuous data output, eg, for periodically transmitting weight data to a printer or continuously to a remote display. RS-422 and RS-485 are typically used in multidrop systems where a single computer may be used to communicate with several scales. Scales can be part of a local area network such as ARCnet, together with other devices such as printers, controllers, and computers. Scales often have an analogue output such as 4-20mA or 0-10 VDC for communication with controllers such as PLCs. Some weight indicators have digital inputs and outputs (I/O); the inputs, for example, could be used for remote operation of scale buttons such as print or tare, whereas the outputs could be used to activate alarms when a particular weight is achieved. Weight indicators designed specifically for batching applications are also discussed herein.

Tank and Hopper Scale Design Considerations

Industrial applications often require that bulk materials or liquids be weighed in hoppers, silos, tanks, or reactor vessels, referred to collectively as vessels. Because they come in such a wide variety of sizes, shapes, and capacities, scales using these vessels as load receivers are not typically available as standard products. Vessels are usually custom-fabricated to suit a particular application, then mounted on a scale. Some can be mounted on a standard scale such as a bench, portable, or floor scale. More typically, a number of weigh modules are used to support the vessel. This offers the scale designer great flexibility but certain precautions are necessary in order to construct an accurate scale. Some of the more important factors associated with the design of vessel scales are discussed herein.

Weigh Modules. Load cells have a single axis along which load should be applied; forces applied in other directions or torques can have a very detrimental effect on weighing accuracy (13). Weighing technology is often applied to existing vessels which were not designed with weighing in mind.

Complicating factors can include less-than-ideal sites, excessive deflections of the vessel and support structure, and thermal expansion and contraction of the vessel, particularly where heated materials are weighed. All of these factors can contribute to poor scale performance by applying extraneous forces to the load cells. To accommodate these problems, load cells are incorporated into weigh modules, which are mechanical assemblies designed to introduce the load correctly to the cell while allowing the vessel to expand and contract, and deflect under load. Weigh modules typically have top and bottom plates that allow them to be inserted conveniently between the vessel and support structure. Several weigh modules are usually used per scale to provide the required stability. Weigh modules are available in capacities of from 25 kg to 100 t or more; they use both hydraulic and strain-gauge load cells, the latter being the more common. The outputs from the various load cells are summed and a signal provided to an indicator for display and control.

Tension and Compression Arrangements. Vessel weighing configurations can be divided into tension and compression systems. Tension systems are often used where a suitable overhead structure already exists, or where the area under the vessel must be kept clear of obstruction. Figure 17 illustrates the use of tension weigh modules on a weigh hopper. Each module typically consists of an S-type load cell, two rod end ball joints, and two clevises. This arrangement allows the load to be applied correctly in the axial direction, while minimizing any bending or twisting forces. Tension systems easily accommodate thermal expansion or contraction of the vessel because of the long suspension rods and hardware used. The hardware is simple and the floor space under the vessel remains free of obstruction. Disadvantages include the need for a rigid overhead structure, and a relatively limited capacity, usually 20 t or less. Lateral checking may be required to prevent swaying or twisting.

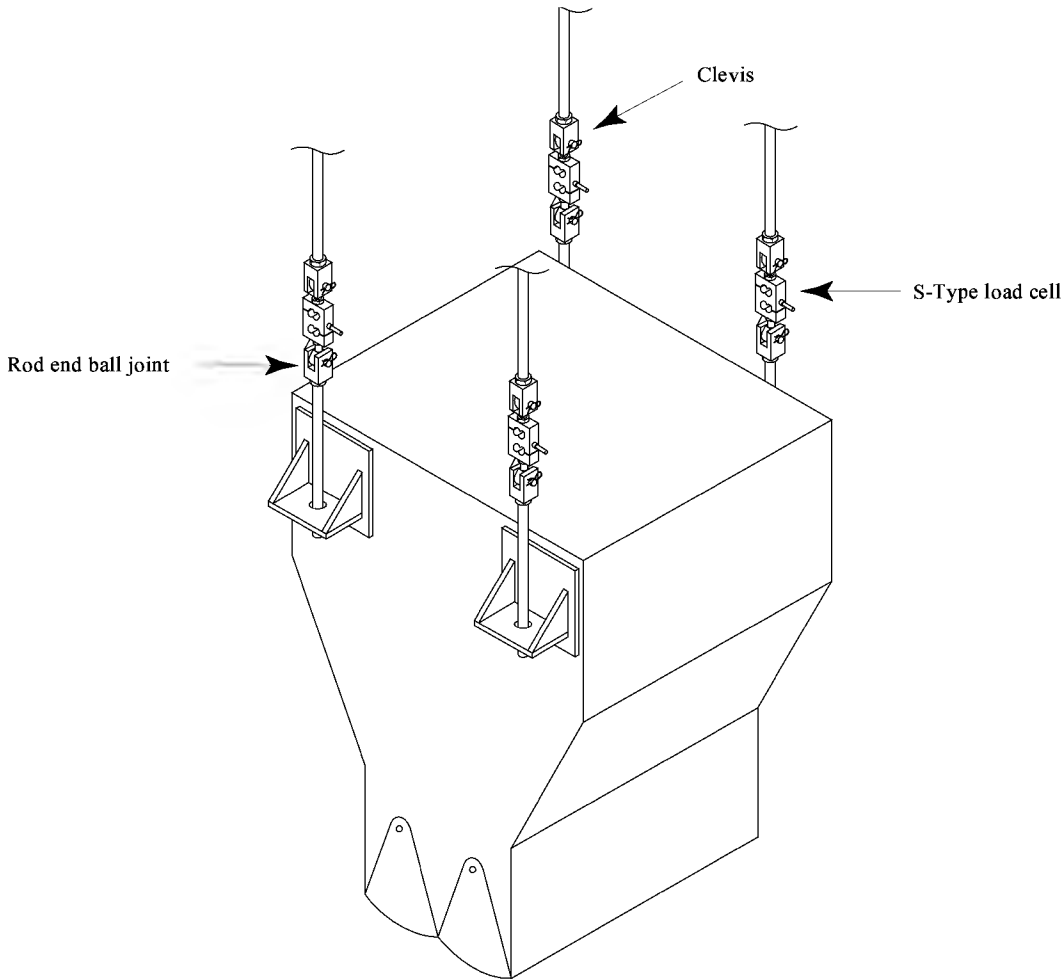


Fig. 17. Tension configuration.

Figure 18 illustrates a vessel mounted on compression weigh modules. Compression systems can accommodate a wide variety of vessel configurations and support structures, including concrete slab and structural steelwork. Weigh modules with capacities from 25 kg to 100 t and more are available. Most designs are self-checking and some include protection against tipping; both make the installation easier. On the other hand, care must be taken to level and align the modules carefully, and the floor-level modules can be subject to flooding, severe washdown, and debris accumulation, which can adversely affect accuracy and cause corrosion.

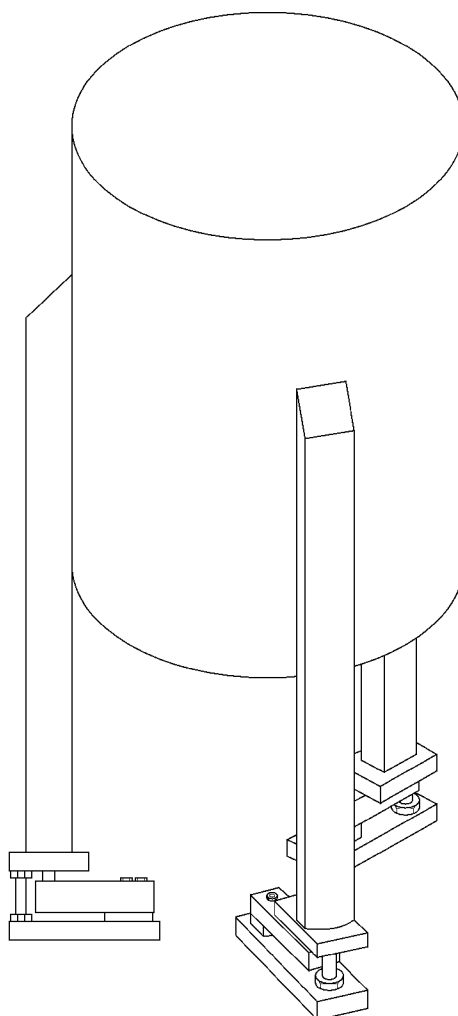


Fig. 18. Compression configuration.

Weigh Module Capacity. The number of weigh modules to be used is often dictated by the geometry of the vessel. Low capacity tanks can be hung from a single load cell, but most vessels are supported by three or four modules. In the case of horizontal tanks or upright vessels of square or rectangular cross section, it is often most convenient to use four weigh modules. In the case of vertical vessels of circular cross section, it is common to use three modules symmetrically arranged. Three-module arrangements are preferred; when more than three are used, it is usually necessary to shim some of the modules to achieve correct load distribution. Systems using more than 8 modules are rare for this reason.

Weigh module capacity can be calculated from the following:

$$\text{weigh module capacity} = \frac{K(\text{live load} + \text{dead load})}{N}$$

where K is a safety factor usually between 1.25 and 2, and N is the number of supports, usually the same as the number of weigh modules. The live load is the total weight of material to be weighed on the scale; it should be conservative and take into account variations in material density and moisture content. The dead load is the total weight of the empty vessel including all equipment mounted on the vessel, heating and cooling fluids in a jacketed vessel, the weight of piping supported by the vessel, etc. The factor K is to some extent dictated by the capacities of available weigh modules. If the scale capacity and dead load are known with certainty, and the scale will be loaded without shock, it may be acceptable to use $K = 1.25$. Where the capacity and dead load are based on estimates, where the vessel is subjected to vibration or material surges, or where the weigh modules are not loaded equally, then K values closer to 2 should be used. If the material is in large chunks and falls from a height onto the scale, there is the danger of shock damage to the load cells, and a K factor larger than 2 may be necessary. It must be noted that all weigh modules on a particular vessel must be of the same capacity.

Mechanical Considerations. The support structure for a weigh vessel needs to be rigid so as to minimize undesirable forces on the weigh modules (14). This rigidity also has the beneficial effect of increasing the natural frequency of the scale, which is important where accuracy and speed are required. Support points need to be equally rigid to prevent the transfer of load between weigh modules; also, vessels with long legs need bracing to prevent lateral deflection of the legs as the load is applied. Where modules are used, as shown in Figure 17, gussets should be used to support the horizontal bracket; the vessel wall may need reinforcement also. For highest weighing accuracy, the vessel should be located where it is free of vibrations from rotating equipment or vehicular traffic. Where more than one vessel is supported on a common structure, care must be taken to avoid deflection of one vessel as another is being filled or emptied.

As a vessel is loaded, it moves downward because of deflection of the load cells and support structure. Pipes rigidly attached to a vessel restrict its free movement and assume some portion of the load that cannot be measured by the load cells. This is very detrimental to scale accuracy. Deflection of the load cell is unavoidable; deflection of the vessel support structure should be minimized. Anything which increases vessel deflection, eg, rubber pads used for shock protection, must be avoided. The total number of pipes should be minimized and be of the smallest diameter, thinnest wall possible. Pipe runs to weigh vessels must be horizontal and the first pipe support should be as far as possible from the vessel. Alternatively, a section of rubber hose or flexible bellows should be used to make the final connection to the vessel. The scale should be calibrated using weights, not by means of an electrical simulation method, which cannot account for the effects of the piping or test the correct functioning of the scale.

If a weigh vessel is located outdoors, the effect of wind forces on the vessel's stability should be considered, particularly if it is a tall slender vessel in an exposed location. An empty vessel may be in danger of tipping; wind forces transfer weight from one module to another and, when full, may cause load-cell damage. In many areas the effects of earthquakes must be considered in the design of the scale and its supporting structure. Building codes exist for the design of structures subjected to wind or seismic loading (15,16). In many instances a weigh module can be selected which can withstand these effects; however, in some cases it may be necessary to add additional horizontal or vertical bracing. Vessel stability can be greatly improved if compression mounts are applied in a horizontal plane close to the vessel's center of gravity; this arrangement is convenient if the vessel passes through a floor, for example. All vessels should have safety supports that can hold the vessel if failure of the primary support could lead to loss or injury. This safety backup could be provided by loosely fitting chains or check rods.

It may be necessary to contain dust by enclosing a weigh hopper and using dust seals or flexible connections to seal openings. Figure 19 shows an arrangement where the top of the hopper is fixed to the structure, and the hopper must have an effective vent which minimizes even transient pressure

surges; otherwise, unwanted vertical forces will be applied to the scale.

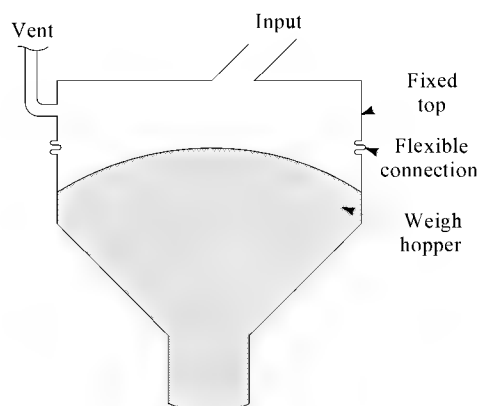


Fig. 19. Hopper top fixed to structure.

Figure 20 shows an arrangement which is unaffected by air pressure fluctuations, because any force applied to the material is canceled by an equal and opposite force applied to the inside top surface of the hopper. It may be desirable or necessary to vent the hopper for efficient material handling.

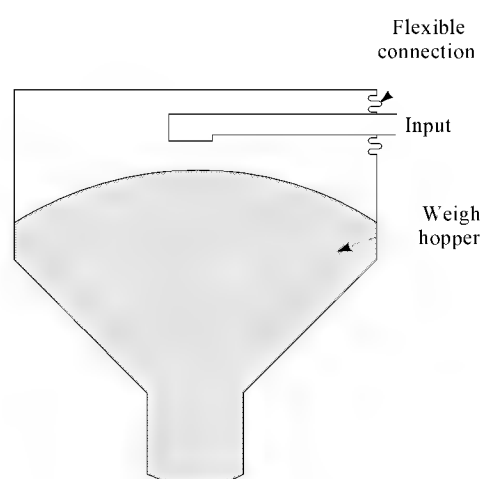


Fig. 20. Top attached to hopper.

Calibration. The greatest accuracy will be achieved by calibrating with test weights equivalent to scale capacity. Provision should be made for test weights when designing a weigh vessel by providing, for example, a shelf on which the weights can be placed, or eyes from which the weights can be hung. Scales used in commercial applications must be calibrated using test weights or material substitution methods using a specified minimum amount of test weights (5). The substitution method can be used where the amount of test weights available is small compared to scale capacity. The weights (15% of scale capacity, for example) are used to calibrate the scale roughly initially and the scale reading is noted; the weights are removed and any available material is then added until the scale reads the noted value. The test weights are added to the scale along with the material added previously and the new reading is noted; the weights are then removed and more material added until the scale reads the second value noted. This process is repeated until a load close to scale capacity has been added and the final calibration can be performed with what is now a known quantity of material and test weights.

Calibration can also be accomplished using material weighed on another scale. The accuracy of this method depends on the accuracy of the other scale, and care must be taken not to lose any of the weighed material. Scales can also be calibrated electrically using a load cell simulator if the load cells' rated outputs are known accurately. This method does not test the mechanical functioning of the scale and is not very accurate, particularly if it has attached piping that restricts its vertical movement.

Feeding Equipment. For consistency in filling or proportioning operations, the design of the feeding equipment is critical. A front-end loader can be used to fill a hopper scale; however, it will be very difficult to achieve a specific weight repeatedly. If a belt-conveyor is used for filling, the results can vary because of lumps and variations in the quantity of material on the belt. The problem will be compounded if the conveyor coasts after power is removed. The scale and material-handling equipment must be designed as a system to achieve the required accuracy.

The choice of feeder is often dictated by the material properties. For free-flowing materials that do not tend to bridge, rathole, or flood, electric vibrating feeders, belt feeders, screw feeders, and power-operated gates are used. For material that is predominantly in large lumps, recommended types include electric vibrating feeders, belt or slat feeders, and screw feeders of proper design if lumps are not too large. For finely pulverized materials that tend to flood, screw feeders are recommended, preferably double-flight or half-pitch (an auxiliary gate may be required), and rotary vane or star feeders. For some applications it may be desirable to use a rotary vane feeder in conjunction with a vibrating feeder; this combination prevents flooding, and the desired smooth control of flow is achieved because of the vibrating feeder.

The highest throughput can be achieved by filling a hopper directly from a silo or upper surge hopper (Fig. 13) equipped with a feed gate. An upper surge hopper is also useful for eliminating the erratic flow associated with devices such as belt-conveyors; it also allows the incoming material to be conveyed continuously despite the batch nature of the weighing process.

Where the scale activates an output, indicating that the desired weight has been achieved, the feeding equipment must stop operating immediately. Some feed devices may need a brake to eliminate coasting. Even if a feeder stops instantly, there will be a certain amount of *in-flight* material added to the scale after the desired weight has been achieved. The average amount of in-flight material can be determined and the scale can be programmed to *pre-act*, ie, it activates the stop output signal when the weight on the scale has reached the desired weight minus the in-flight material. For this to be effective, the rate of material flow must be consistent. Best results are achieved by minimizing the amount of in-flight material; hence the height of the feeder over the scale should be minimized and the rate of fill should be very slow. Where the latter is not practical, a *two-speed fill* can be used. Figure 21 shows a storage hopper with a three-position simplex gate. At one extreme it is fully open, allowing fast (*bulk*) fill; at the other extreme it is fully closed, and in the intermediate position (illustrated) it feeds slowly (*dribble*) through a small notch in the gate. Figure 22 illustrates a similar arrangement for filling liquids; in this case, bulk fill is achieved by opening both solenoid valves, and dribble fill is achieved by closing the larger one. Motor-driven feeders can perform bulk and dribble fill conveniently by varying the motor speed. The ratio of bulk to dribble volume can be as much as 1,000 to 1; dribble is often used for the last 5–10% of the material filled. A further refinement is an *autojog* feature. A pre-act is set to ensure that the filled weight is always light; after filling and settling, the weight is

compared to the required value. If it is low, the scale jogs the feeder until the batch is within tolerance.

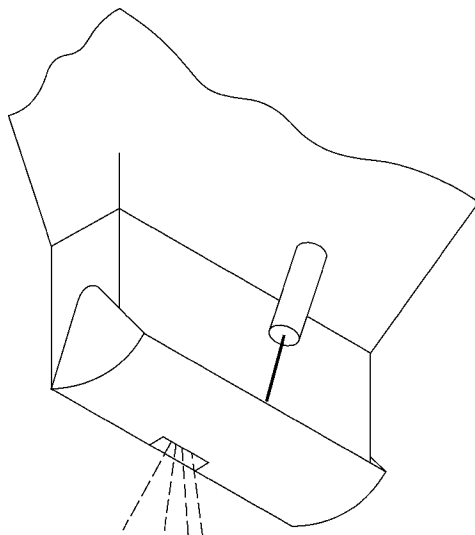


Fig. 21. Hopper with bulk and dribble fill.

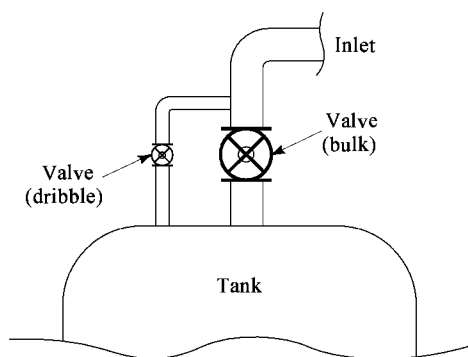


Fig. 22. Tank with bulk and dribble fill.

Discharging the Scale. Where a gate is used to discharge a hopper scale, very little control can be exercised over the rate of discharge. In some cases, particularly where the various scales of a simultaneous batching system discharge onto a single conveyor belt, it is desirable to control the discharge rate so as to premix the materials. This can be accomplished by using a suitable feeding device in place of the gate. Also, using a controllable feeder as the discharge device, it is not necessary to empty the hopper or tank completely since material can be accurately weighed out of the hopper, as well as into it. In the case of materials that cling to the sides of a hopper, use of this device provides a definite advantage. In the case of liquids, more rapid discharge of a definite amount can be accomplished because the last material does not have to be discharged fully with diminishing pressure head; this method is also very useful in the case of very viscous liquids. Of course, it cannot be used when different materials are accumulated in a scale hopper because the composition of the material remaining in the hopper would vary.

Indicators and Controllers. A batching scale can be controlled manually by an operator observing the weight display and operating the feed and discharge devices at the appropriate time. Some scale indicators are designed specifically for automatic control of batching systems, and the recipe and batching routine can be programmed at the indicator. These indicators can perform functions such as two-speed fill, pre-act, and autojog. They also have digital I/O capability; the inputs can, for example, be used to start and stop the batching process, whereas the outputs are used to control the filling and discharge equipment. This system works well in automating a standalone batching system, because fast reaction times and repeatability can be achieved. The disadvantages are that it is difficult to integrate into a larger system and that the programming can be slow and error-prone.

Another approach to the control of batching systems is to use a relatively simple scale indicator which does nothing but supply weight data to a controller such as a PLC, which in turn controls the fill and discharge system. The communications between the scale and controller can be serial such as RS-232, or it can be an analogue signal such as 4-20 mA. While this system minimizes operator interaction with the scale, it sacrifices speed and accuracy because of the relatively slow communications.

Yet another approach is a hybrid of the last two, and is particularly applicable to complex processes. The overall process, including the batching scale, is controlled by a controller, such as a PLC. The PLC downloads the recipe to the scale and tells it when to commence; the batching is then controlled completely by the scale through its own digital I/O. The advantage is that the scale is fully integrated into the larger process without compromising speed and accuracy of the batching operation.

Regardless of which form of control is used, the update rate of the indicator is critical. The update rate is the number of times per second the scale samples the weight on the load cells and compares the result with the programmed target weight. Between updates, the indicator is blind to the material being added to the scale. For conventional weighing scales the update rate might be no more than 5 or 10 updates per second, because time is not critical. In high speed filling, a lot of material can be filled in 1/10 of a second, and weight indicators with 20 to 100 updates per second are typically used.

Another important feature of the scale is the filtering. The weight signal from the load cells can be contaminated by noise generated by vibrations from adjacent machinery, rotating equipment mounted on the scale, or vehicular traffic. Noise can also be electrical, eg, electromagnetic interference (EMI) or radio frequency interference (RFI) from adjacent electrical equipment. To determine the actual weight from this signal, sophisticated batching indicators use a combination of analogue and any of several types of digital filtering that can be optimized for a given scale. The disadvantage of too much filtering, or filtering of the wrong type, is that it increases the reaction time of the indicator.

BIBLIOGRAPHY

"Weighing and Proportioning" in *ECT* 1st ed., Vol. 15, pp. 25-34, by D. B. Kendall, Toledo Scale Company; in *ECT* 2nd ed., Vol. 22, pp. 220-241, by D. B. Kendall, Toledo Scale Company; in *ECT* 3rd ed., Vol. 24, pp. 482-501.

1. P. Giesecke and T. Preusser, *Of Scales and Scale Builders*, Carl Schenck AG, Darmstadt, Germany, 1991, pp. 7, 44.
2. J. Ayto, *Dictionary of Word Origins*, Arcade Publishing Inc., New York, 1990, pp. 49, 459.

3.

E. O. Doebelin, *Measurement Systems Application and Design*, McGraw-Hill Kogakusha, Ltd., Tokyo, 1975, p. 605.

4.

W. H. Hayt, Jr., *Engineering Electromagnetics*, 4th ed., McGraw-Hill Book Co., Inc., New York, 1981, pp. 293–345.

5.

Handbook 44, Specifications, Tolerances, and Other Technical Requirements For Weighing and Measuring Devices, National Institute of Standards and Technology, Gaithersburg, Md., 1996.

6.

International Vocabulary of Basic and General Terms in Metrology, 2nd ed., Organisation Internationale de Metrologie Legale, Paris, 1993.

7.

Glossary of Weighing Terms, A Practical Guide to the Terminology of Weighing, Mettler-Toledo Inc., Worthington, Ohio, 1994.

8.

OIML R76-1, *Non Automatic Weighing Instruments*, Organisation Internationale De Metrologie Legale, Paris, 1992.

9.

ASTM E617, *Standard Specification for Laboratory Weights and Precision Mass Standards*, American Society for Testing and Materials, Philadelphia, Pa., 1991, p. 2.

10.

J. L. Meriam, *Statics*, 2nd ed., John Wiley & Sons, Inc., New York, 1975, pp. 215–217.

11.

G. B. Anderson and R. C. Raybold, *Tech Note 436*, National Institute of Standards and Technology, Gaithersburg, Md., 1969.

12.

G. F. C. Rogers and Y. R. Mayhew, *Engineering Thermodynamics, Work and Heat Transfer*, 2nd ed., Longman Group, London, 1967, p. 551.

13.

A. Goh, *Chem. Eng.*, 47–51 (Dec. 1996).

14.

Do-It-Yourself Guide to Building Tank Scales, Mettler-Toledo Inc., Worthington, Ohio, pp. 34–35, 1997.

15.

Uniform Building Code, International Conference of Building Inspectors, Whittier, Calif., 1994.

16.

ASCE 7-93, Minimum Design Loads for Buildings and Other Structures, American Society of Chemical Engineers, New York, 1993.

General References

Handbook 105-1, Specifications and Tolerances for Reference Standards and Field Standard Weights and Measures, National Institute of Standards and Technology, Gaithersburg, Md., 1990.

A Procedure for the Specification, Calibration and Testing of Strain Gage Load Cells for Industrial Process Weighing and Force Measurement, Institute of Measurement and Control, London.

W. J. Youden, *Experimentation and Measurement (Publication 672)*, National Institute of Standards and Technology, Gaithersburg, Md., 1991.

Technical Publication, National Industrial Scale Association, Birmingham, Ala.

W&M Today, *Newsletter of the NCWM*, National Institute of Standards and Technology, Gaithersburg, Md.

OIML Bulletin, Organisation Internationale De Metrologie Legale, Paris.

H. Colijn, *Weighing and Proportioning of Bulk Solids*, Trans Tech Publications, Aedermannsdorf, Switzerland, 1975.

Weights and Measures 24-hour fax line (800-925-2453), National Institute of Standards and Technology, Gaithersburg, Md.

Weighing and Measurement, Key Markets Publishing Co., Rockford, Ill., published bimonthly.

Sensors, Helmers Publishing Inc., Peterborough, N.H., published monthly.

I&CS (Instrumentation and Control Systems), Chilton Company, Radnor, Pa., published monthly.

Waegen + Dosieren, Verlagsgesellschaft Keppler-Kirchheim mbH, Mainz, Germany, published bi-monthly.

Chemical Processing, Putman Publishing Co., Chicago, Ill., published monthly

Tom Leahy
Mettler-Toledo, Inc.

WELDING

Welding comprises a group of processes whereby the localized coalescence of materials is achieved through application of heat and or pressure (1–3). It represents a fabrication technology of extremely broad scope; welding principles are used in nearly every product fabricated, including metals and, increasingly, nonmetals. Obvious examples are bridges, buildings, ships, space vehicles, automobiles, offshore platforms, pipelines, pressure vessels, and consumer appliances; less evident examples include plastic cases for computer disks and the very large number of miniature welds used in microcircuits. This field of great technical complexity utilizes robots, computer-controlled machines, and high concentrations of energy in the form of plasma arcs, lasers, and electron beams. Welding involves very rapid processes, including heat transfer, chemical reactions, and metallurgical reactions, and requires complex methods of analysis to control stress, distortion, and fracture.

The principal welding processes are of comparatively recent origin. Forge welding, soldering, and brazing can be traced to ancient times, but the modern processes of arc welding, resistance welding, and gas welding (as well as thermit welding) were discovered in the 1880s. The use of welding in production and repair increased slowly through the early 1900s and received a boost by successes in emergency ship repair during World War I. Welding saw increased use in building construction during the 1930s, and new welding processes, eg, submerged arc welding, were discovered. World War II provided tremendous impetus for further development, including spectacular uses in all-welded ships and for military tanks and development of new processes for welding aluminum. Since the 1940s many new welding techniques have appeared, including electron-beam, laser, and ultrasonic processes (4).

Thus, from the origins of modern welding in the 1880s many new processes have evolved, and today the American Welding Society recognizes nearly seventy methods of welding and more than twenty allied processes employing thermal cutting, thermal spraying, and adhesive bonding (1), some of which are identified in Table 1.

Table 1. Welding Processes^a

Process	Abbreviation	Process	Abbreviation
arc welding		resistance welding	RW
	SMAW	resistance spot welding	RSW
shielded-metal arc welding		resistance seam welding	RSEW
	GTAW	projection welding	RPW
gas–tungsten arc welding		flash welding	FW
plasma arc welding	PAW	solid-state welding	SSW
gas–metal arc welding	GMAW	diffusion welding	DFW
flux-cored arc welding	FCAW	explosion welding	EXW
stud arc welding	SW	forge welding	FOW
submerged arc welding	SAW	friction welding	FRW
oxyfuel–gas welding	OFW	ultrasonic welding	USW
oxyacetylene welding	OAW	other welding processes	
oxyhydrogen welding	OHW	electron beam welding	EBW
pressure-gas welding	PGW	laser beam welding	LBW
brazing	B	electroslag welding	ESW
resistance brazing	RB	thermit welding	TW
furnace brazing	FB	allied processes	
induction brazing	IB	thermal spraying	THS
soldering	S	adhesive bonding	ABD
dip soldering	DS	thermal cutting	TC
wave soldering	WS		
torch soldering	TS		

^a Ref. 1.

Arc-Welding Processes

In arc welding, the coalescence of metals is achieved through the intense heat of an electric arc, which is established between the base metal and an electrode. The processes listed in Table 1 are differentiated by various means of shielding the arc from the atmosphere (1–3).

In arc processes, either d-c or a-c current is used to establish and maintain an arc between the base metal and an electrode. The electrode itself may be consumed by melting and thus become part of the weld, acting as a filler metal, or it may be a nonconsumable material, eg, tungsten. In the latter case, the heat of the electric arc may simply be used to fuse adjacent base metal (autogenous welding), or a separate filler metal may be added. It is essential that the molten pool of material under the arc, as well as the adjacent solidified but still high temperature metal, be protected from oxygen, nitrogen, and other elements of the atmosphere, since these react with the metal to form oxides and other products that reduce the strength and toughness of a weld. Consequently, various forms of shielding are provided around the arc in the different processes.

Shielded-Metal Arc Welding. The essential features of the SMAW process, commonly called stick-electrode welding, are shown in Figure 1a. The arc acts between the consumable electrode wire and the base metal, and droplets of molten filler metal are transferred to the weld pool. The unique feature of this widely used process is the role of the electrode coating. This coating is decomposed by the heat of the arc and provides both a necessary shielding atmosphere for the arc as well as a slag coating over the weld metal, thus affording still further protection from the atmosphere.

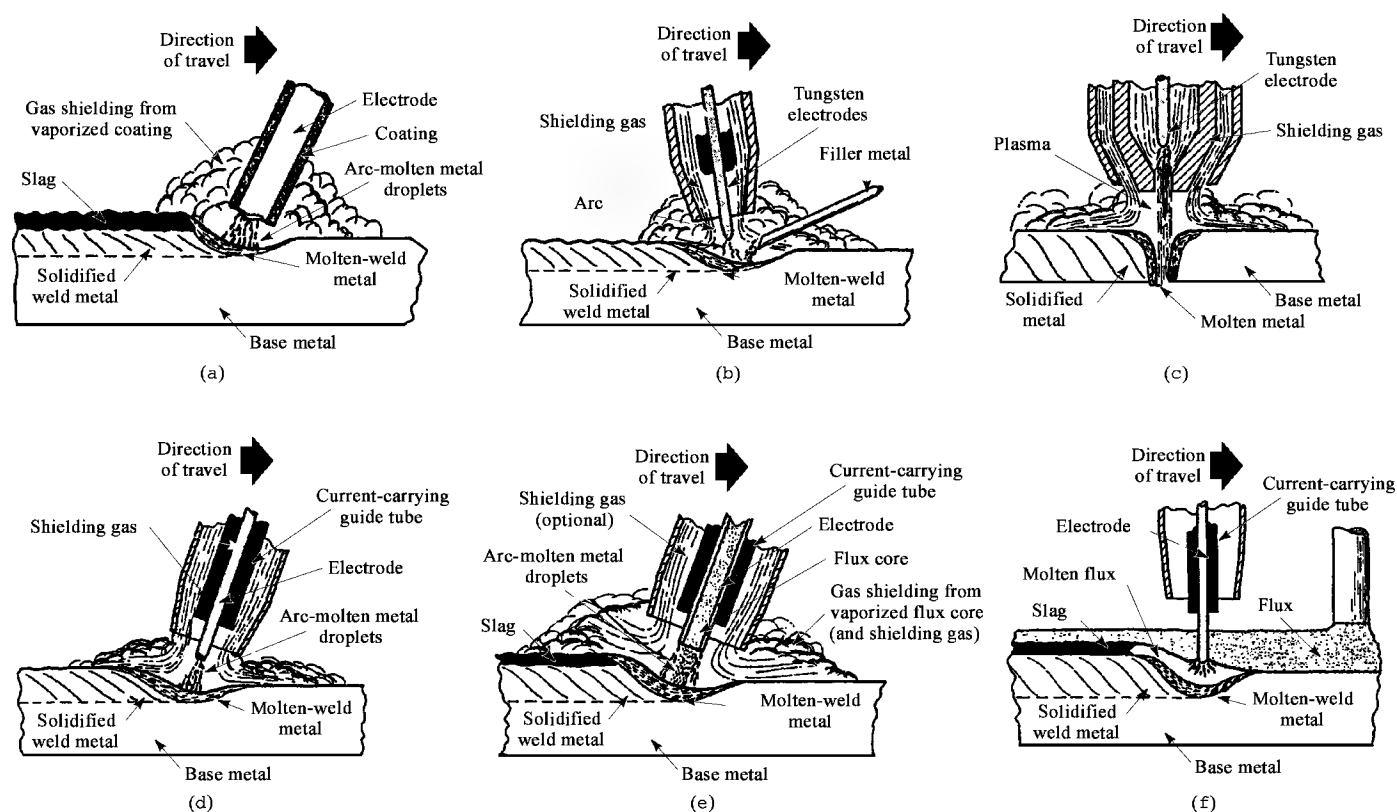


Fig. 1. (a) The shielded-metal arc-welding (SMAW) process. (b) The gas-tungsten arc-welding (GTAW) process. (c) The plasma arc-welding (PAW) process. (d) The gas-metal arc-welding (GMAW) process. (e) The flux-cored arc-welding (FCAW) process. (f) The submerged arc-welding (SAW) process.

In addition to these functions, the coating introduces fluxing agents to the weld pool, assists in establishing the electrical characteristics of the arc, and can be used to provide additional filler metal to the weld. Over a dozen different materials may be used in a single electrode covering, depending on the desired welding characteristics. Common ingredients include cellulose and calcium carbonate to provide shielding gas, titanium dioxide and silica to provide slag, sodium silicate to act as a binder, and ferrosilicon to act as a deoxidizer.

Deposition rates using this process are limited by the fact that each electrode contains a finite amount of filler metal. The time required both to change electrodes and to remove the slag coating between each weld pass lowers the overall productivity of the process.

Gas-Tungsten Arc Welding. The GTAW process, often called the tungsten-inert gas or TIG process, is shown in Figure 1b. Here the electrode is nonconsumable and shielding is provided by the flow of inert gas through the welding-torch nozzle. Argon and helium are commonly used, as well as argon-helium mixtures; traces of hydrogen are sometimes present. The arc may be used alone to fuse base metal together or, as shown in Figure 1b, separately added filler rod can be fed into the weld pool.

Plasma Arc Welding. In the transferred-arc mode of the PAW process, shown in Figure 1c, the arc is between a nonconsumable electrode and the base metal, in a manner similar to the GTAW process. The unique feature is the flow of inert gas around the electrode and through a restricted orifice, which constricts the arc to form a plasma jet. A second, outer stream of shielding gas protects the molten metal from atmospheric contamination. In the nontransferred arc mode of the PAW process, the arc is between the electrode and the constricting orifice. This mode is used for plasma spraying and for heating nonmetals.

Gas-Metal Arc Welding. The features of the GMAW process, also called the metal-inert gas or MIG process, are shown in Figure 1d. A consumable bare-wire electrode is fed continuously through the welding torch, and a flow of inert gas through the nozzle provides shielding. This continuous-wire feed leads to higher deposition rates compared to the SMAW process. Common shielding gases include 100% argon, 75–80% argon with the remainder carbon dioxide, and 100% carbon dioxide.

Flux-Cored Arc Welding. The FCAW process is illustrated in Figure 1e. The characteristic feature is the hollow, flux-filled, consumable electrode, which is fed continuously through the welding torch. The decomposition of the flux provides both arc shielding and a slag blanket over the weld. In a variation of the process, a shielding gas flows through the welding nozzle to provide additional protection for the arc.

Submerged-Arc Welding. In the SAW process, shown in Figure 1f, a loose flux is blanketed over the region of the arc. A consumable, bare-wire electrode is fed continuously through the torch into the weld. The molten- and granular-flux blankets provide the necessary arc shielding and slag cover for the solidified weld. This process, often used with operating currents of over 1000 A, provides the highest deposition rate of all the arc-welding processes. The presence of a loose flux makes vertical and overhead welding impractical with the SAW process.

Arc-Welding Systems

The various welding processes result in systems of varying complexity. They include at least the electrode and a device for holding or feeding it, the work piece, the power source, and heavy-duty cabling to provide a complete electrical circuit. Provisions for supply and control of gas and control of wire feed and movement of the electrode assembly are required, depending on process type and degree of automation.

Welding systems are generally classified as manual, semiautomatic, mechanized, and automatic (2). In manual welding, the operator must maintain the arc, feed in filler metal, and provide travel and guidance along the joint. In semiautomatic welding, the welding machine maintains the arc and feeds filler metal, and the operator controls joint travel and provides guidance. In mechanized welding, the welding machine maintains the arc, feeds filler metal, and moves a mounted torch along the joint; the operator manually adjusts the welding parameters based on observation of the process. In automatic welding, the machine assumes all of the preceding functions. Automatic processes may be further classified, depending on the degree of feedback control used in controlling the welding variables. Automated systems may be dedicated to a specific type of production, or they may be flexible, programmable robot systems.

The power source is the core of any welding system. Electrical power is provided by direct-line power or a generator; the latter is driven by an engine or an electric motor. In the case of welding with a-c current, the welding power source may be in the form of a transformer working from line power, or it may be an engine-driven a-c generator. Welding with d-c power may require a transformer-rectifier working from single- or three-phase line current, or a motor-driven generator. Welding power sources are further differentiated by controls imposed on the output current or voltage. One type of power source, used primarily for manual welding, is capable of providing essentially constant current. Another type, used for semiautomatic and automatic processes, is capable of providing essentially constant voltage.

The introduction of inverter power sources with solid-state electronic components has resulted in weight and size reductions of up to 75%.

compared to traditional power supplies. The efficiency and performance of these power supplies is also increased. Variable-frequency, pulsed-current power supplies for GMAW and GTAW, wherein a base current level pulses to high currents, depend on inverter technology. Pulsed GMAW, with pulsing rates of 60 to 200 times per second, provides spray metal transfer at lower average currents and with less fume and spatter; the ability to weld thinner material and to make out-of-position welds is also increased. Pulsed GTAW, with pulsing rates of 1 to 20 times per second, provides greater penetration for a given average current and provides for thin-section welding at lower average currents. High (ca 20 kHz) frequency, pulsed GTAW can be used to provide a stiffer, more directional arc.

The system for shielded-metal arc welding, shown in Figure 2a, is the simplest system. It consists of the power source, electrode and holder, the base metal, and the electrical cables or leads. When the arc is struck, a complete electrical circuit is provided. With d-c welding, the electrode may be either negative (straight polarity) or positive (reverse polarity). Shielded metal arc welding is only used manually.

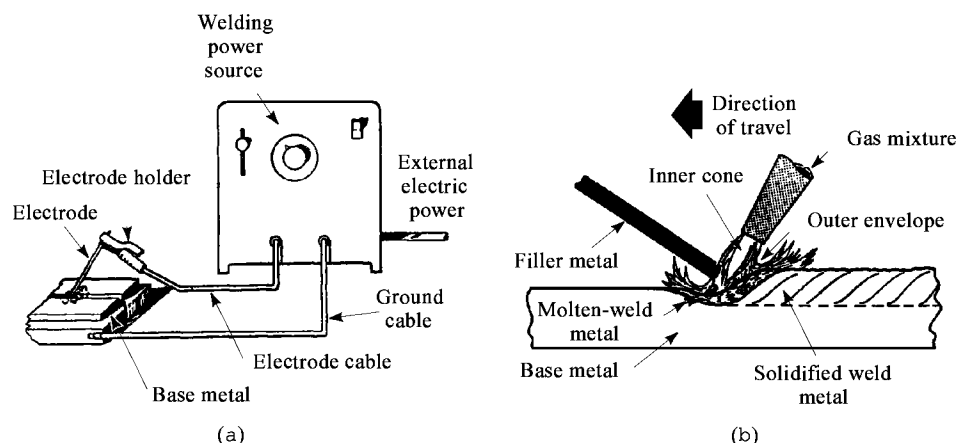


Fig. 2. (a) The shielded-metal arc-welding system. (b) The oxyacetylene welding process.

A gas-tungsten arc-welding system is more complex. In addition to the components of the shielded-metal arc system, provisions must be made for the inert gas supply and water or air cooling of the welding torch. GTAW systems may range from manual to automatic.

A semiautomatic flux-cored arc system is shown in Figure 3. Controls and drive motor are required for continuous feed of the welding wire to the torch. If a shielding gas is used, provisions for control of this gas supply are needed, and the torch configuration is different. The flux-cored system may be used for gas-metal arc welding (GMAW) by using the shielding gas and the welding torch configuration for use with gas. Both FCAW and GMAW systems can be used in automatic operations.

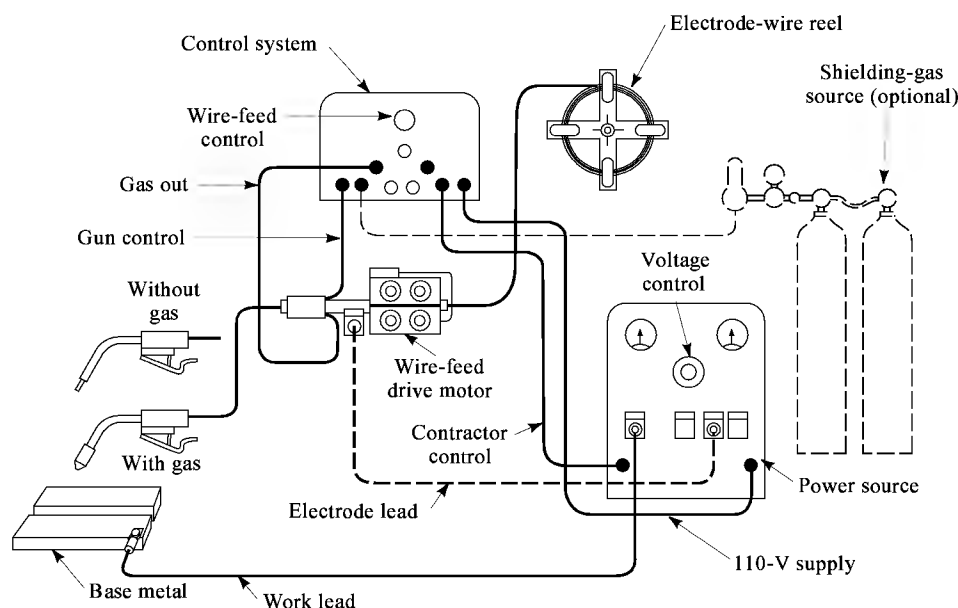


Fig. 3. The flux-cored arc-welding (FCAW) system (2).

A mechanized, submerged-arc-welding system requires a power source, control system, and wire-feed provisions. Granulated flux is fed into the weld joint from a flux hopper, which travels ahead of the welding arc as part of the electrode carriage. Control of the welding torch along the seam of the base metal can be accomplished by a motorized electrode carriage moving along a straight beam, or by a positioner capable of placing the welding torch at any position over a base metal that is itself continually changing position.

Other Important Fusion Welding Processes

Oxyfuel–Gas Welding. This process, commonly called gas welding, uses the heat of combusting gases to melt and coalesce base metals. Although several different fuel gases, eg, propylene, hydrogen, MAPP, or methane, can be added to the oxygen, the oxyacetylene flame is the most widely used because its high (3100°C) flame temperature is needed to weld steel. The essential features of the process are shown in Figure 2b. The heat of the inner cone of the flame melts both the base and the filler metal. The overall system is simple; it consists of gas regulators, hoses, the welding torch, a high pressure oxygen tank, and the acetylene tank, which contains liquid acetone into which the acetylene is dissolved. Gas welding, once the most widely used welding method, has been supplanted by arc processes in commercial manufacture. However, it is still used in small welding shops, for repairs, and by the home craftsman.

Resistance Welding. As noted in Table 1, resistance welding comprises several processes; the most widely used is resistance spot welding (RSW). The principles of RSW are quite different from those underlying the processes previously described (Fig. 4). The workpieces are firmly clamped between copper electrodes, and an electric current is passed through the assembly. Heat is generated by the electrical resistance of the components, the maximum heat occurring at the interface between the workpieces. A nugget of metal at the interface region is melted, at which time the current is shut off and the clamping force on the electrodes released. The entire sequence typically requires less than a second. Resistance spot welding is widely used in the

production of thin-gauge, sheet-metal assemblies, eg, automobile bodies. It is a high speed production process and is suitable for automation. Although mainly used for welding steel, it is also used with aluminum and other materials (5).

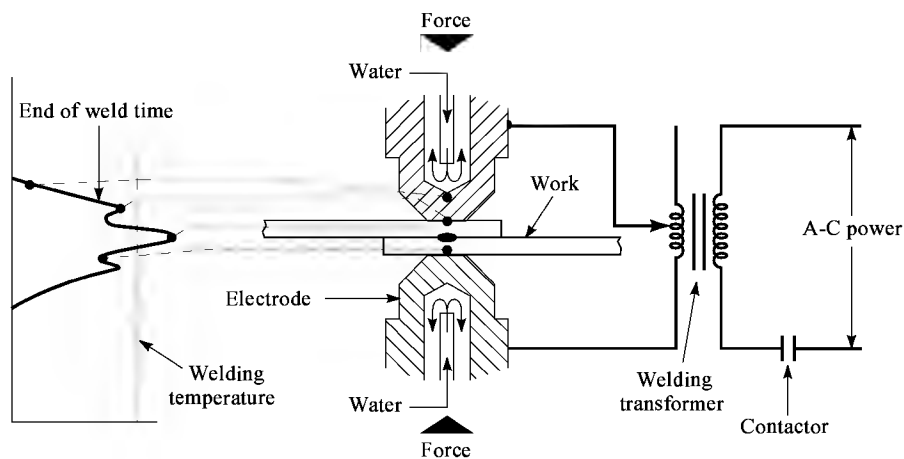


Fig. 4. The resistance spot-welding process (2).

Electroslag Welding. In this process, the heat of a molten slag coalesces the base and filler metals. Electric current flows from a consumable electrode through a molten metallurgical slag into the molten weld metal. The electric resistance of the slag provides the heat to maintain the slag in a molten state and to melt the weld and base metals. In order to contain the molten metal and slag in the gap between the vertical base-metal plates, water-cooled copper shoes are placed on each side of the gap. The molten metal progressively solidifies upward from the bottom of the pool, welding the plates together in one pass. Although initially an electric arc melts the slag, steady electroslag welding is a nonarc process. It has acquired extensive use in welding thick sections of metal. Deposition rates are very high. It is an automatic process that is generally used in the vertical direction.

Electron-Beam Welding. This welding process achieves the heat necessary for coalescence by bombarding the base metals with a concentrated stream of electrons. Electrons emitted by a heated filament are accelerated between cathode and anode and then focused into a narrow beam onto the base metal. The kinetic energy of the electrons is converted to heat and x-rays upon impact. The electron beam permits the welding of thick sections, but with a very narrow zone of molten metal. Electron-beam welding is usually performed in a very high or hard vacuum. Modifications have led to soft and out-of-vacuum welding, although penetration capability is sacrificed (6).

Laser-Beam Welding. The heat of coalescence is produced by focusing the beam of light (photons) from a high power laser on the base metal. Laser welding is considered a high energy-density (ca 10^6 W cm^{-2}) welding process, with weld characteristics similar to the electron beam. Thus, at low energy, surface melting occurs, but at high energy, a deep penetration of the base metal is possible. Unlike the electron beam, the laser does not require a vacuum and can be used to weld nonconductive materials such as polymers. Because the laser beam is essentially a light beam, it is easily controlled by mirrors, beam splitters, and other optical devices, which leads to great flexibility in its use (6). High reflectivity materials such as aluminum are noted for being more difficult to weld using lasers.

Nonfusion Welding Processes

Solid-State Welding. Solid-state welding comprises a group of welding processes wherein a bond is made between two base materials upon the application of pressure at a temperature below the solidus of the base materials (Table 1). Interlayers are sometimes used. By joining materials in the solid state, many of the difficulties of the fusion processes are avoided.

Friction welding is the most common solid-state process. One type, direct-drive friction welding, is accomplished by spinning one piece rapidly against a second, stationary piece to produce frictional heating. When the metal is at a forging temperature below the solidus, rotation is stopped and axial force is applied. Heated metal, together with oxide and surface contamination, is squeezed out of the interface between the two pieces in the form of a metal flash, which can be trimmed to form a smooth surface.

Geometries that have rotational symmetry on at least one of the parts are particularly suitable for friction welding: round bars and tubes, as well as bar-to-sheet and tube-to-sheet applications. The speed of the process, as well as its ease of automation, promote high volume production. Certain metals are not suitable for friction welding, ie, nonforgeable metals, which tend to crumble under heat and pressure, and free-machining alloys, the inclusion content of which can lead to difficulties as the parts are spun and forged together.

Another variant of the friction welding process, linear friction welding, uses servo-hydraulics pumps to vibrate parts back and forth against each other. Bond areas of approximately 1000 mm^2 can be joined; the attachment of turbine blades to rotors is a prevalent application of this technology.

Brazing and Soldering. In brazing and soldering processes, a molten filler metal flows by capillary action into the closely fit joint between the base metals at a temperature below the melting point of the base metal. Bonding is accomplished by metal-to-metal adhesion, which may involve the formation of intermetallic compounds at the interface between the base and filler metals. Although these characteristics are common features of brazing and soldering, the two processes are differentiated by the melting temperature of the filler metal. In brazing processes, the filler metal melts at a temperature above 450°C , whereas in soldering, the melting point of the filler metal is below 450°C (7). Common solder alloys contain lead with amounts of tin varying from less than 5 wt % to around 60 wt %; the eutectic composition of 61.9 wt % tin and 38.1 wt % lead melts at 183°C . Common applications for lead-tin solders including plumbing and electronics. Current (ca 1997) efforts to achieve lead-free solders are concentrating on alloys such as tin with 3.5 wt % silver.

Transient liquid-phase bonding is a technique that can form a brazing or soldering joint at low temperature that can withstand service at higher temperature, sometimes approaching the melting temperature of the base metal. A boron-rich brazing filler paste is used for the repair of gas-turbine engine components. Boron acts as a melting-point depressant and allows the paste to melt and fill defects. The small, highly mobile boron atoms rapidly diffuse into the parent metal at brazing temperatures; the remaining filler metal has a melting point that approaches that of the base metal. Transient liquid-phase soldering for tooling and microelectronics assemblies depends on the dissolution of a high melting-point metal into a solder to form an intermetallic compound which has a higher melting point than the original solder.

Wetting, the ability of the filler metal to spread through the joint, is a critical factor in these processes. Any oxides or other film must be cleaned from the joint area, generally by a fluxing agent, to ensure good wetting. Uniform heating of the joint area on large parts may best be accomplished in a furnace. Brazing can be used to join ceramics as well as metals. Metal-to-ceramic brazing is often accomplished with the addition of reactive metal, such as titanium or zirconium, to the braze material.

Joining of Polymers and Adhesive Bonding

Polymers are characterized as thermosetting and thermoplastic with respect to the methods by which they are joined. Thermosetting polymers are permanently hard and do not soften upon the application of heat; they are joined by mechanical fasteners and adhesives. Several methods have been devised to join thermoplastic polymers, as well as thermoplastic composite materials, which soften upon heating.

Hot Plate, Infrared, and Hot Gas Welding. These processes involve external means to heat thermoplastic polymers to a viscous state in

which the interdiffusion of polymer chain molecules can occur with the application of pressure. In hot plate welding, the two surfaces to be joined are forced against a platen heated to a desired temperature based on the composition of the polymer. After heating occurs, the platen is removed and the two polymer faces are forged together to make a weld. The platen is replaced as a heat source by infrared light in the infrared welding process; the forging step after heating remains the same. In hot gas welding, a stream of heated gas or air is directed at both the joint surfaces and a polymer filler material. In a process similar to the arc welding of metals, the polymer filler metal is fed into the joint to complete the weld.

Friction and Ultrasonic Welding. Both rotational and linear friction can be used to melt the interface between two thermoplastics. The parts are then aligned, and a weld is formed as the interface solidifies. Linear friction welding, also called vibration welding, employs frequencies in the 100–500 Hz range; welding can be accomplished in less than one second but is limited to bond areas approximately 400 mm². Ultrasonic welding of polymers involves oscillations of 10–50 kHz, which are dissipated at the bond line to produce heat through both friction and hysteresis. The surfaces to be joined are held together as the sound energy generated by the welding machine is transferred through the parts at right angles to the contacting surfaces. Like the friction process, ultrasonic welding is limited to smaller part sizes.

Adhesive Bonding. As one of the processes allied to welding, adhesive bonding is used particularly in applications where welding or mechanical fastening is either impossible or undesirable (Table 1). Adhesives, which are derived from polymers, are not structural materials and act only to distribute stress over the bond area; as such, proper joint design and minimization of bondline thickness are essential. The selection of an adhesive should take into account the expected service environment, because temperature and atmosphere can degrade adhesives. Stress mode is also a prime consideration in both joint design and adhesive selection. Pure compression, shear, and tension are preferred over mixed stress modes such as peel or cleavage.

Physics and Metallurgy

In most welding processes, the local regions of the adjoining base metals and any added filler metal are melted and resolidified. These features are present, for example, in the arc, gas, electroslog, electron-beam, and laser-welding processes. The analogy of this action to the casting process is often made, where the molten base and filler metals are cast in the mold formed by the nonmolten parts of the base metals. This analogy is of some use because a range of phase transformations is involved as the alloy system solidifies. However, welding differs from conventional casting by the speed of the solidification process. Only a local region of material is melted in welding, and the surrounding, low temperature base metal acts as a large heat sink, producing rapid heat flow from, and the solidification of, the weld zone. This situation produces a complex metallurgical microstructure and physical properties not typical of casting.

Heat Flow. Certain basic features of welding heat flow are illustrated by Figure 5a, which represents isotherms resulting from a welding heat source moving along the joint between two base metals. The central region, under the heat source, is molten. The surrounding elliptical curves are isotherms, which are clustered toward the direction of travel of the heat source and show the temperature distribution at a given instant during welding. Different welding processes, plate thicknesses, and welding speeds modify the details of the clustering and magnitudes of the isotherms, without changing the principles of the behavior shown.

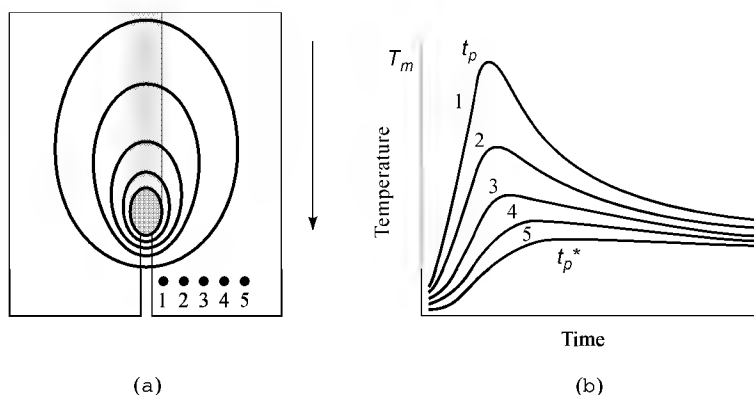


Fig. 5. (a) Temperature isotherms in the region of a moving welding arc. (b) Heat-affected-zone thermal cycles at various locations in the base plate (see text).

An important aspect of welding heat flow is the thermal cycle at a given location in the material. The nature of the cycle depends on the intensity of the heat source, the speed of welding, the thermal characteristics of the material, and the location in the material. The general behavior is shown in Figure 5b. Curve 1, representing point 1 of Figure 5a, is adjacent to the weld and actually reaches the melting point, T_m , as the arc passes. At points 2–5, two characteristics are evident. First, the peak temperatures that occur, plotted along the line t_p to t_p^* , are progressively lower because the heat spreads over larger regions. Second, there is a progressively greater time delay in reaching peak temperature because of the time required for heat conduction. The weld thermal cycle, involving peak temperatures achieved and the speed of heating and cooling, and possibly preheating and postheating, accounts for many of the subsequent complexities of welding metallurgy.

Solidification. The heat of the electric arc melts a portion of the base metal and any added filler metal. The force of the arc produces localized flows within the weld pools, thus providing a stirring effect, which mixes the filler metal and that portion of the melted base metal into a fairly homogeneous weld metal. There is a very rapid transfer of heat away from the weld to the adjacent, low temperature base metal, and solidification begins nearly instantaneously as the welding heat source moves past a given location.

Solidification begins as atoms of weld metal attach themselves to the solid metal grains at the weld pool edge; this initial growth is epitaxial, ie, the atomic orientation of the base-metal grains continues into the weld pool. Weld metal grain growth continues in the direction of the maximum temperature gradient, which is initially perpendicular to the edge of the weld pool and tends to remain so for fast welding speeds, but which turns toward the arc in a direction parallel to the weld axis for a slower welding speeds (Fig. 6). Cellular and cellular dendritic are the most common weld solidification modes; the mode itself is affected by temperature gradient, speed of solidification, and solute content of the weld metal.

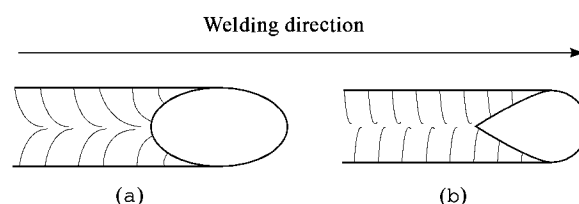


Fig. 6. Weld pool shape and resultant weld-metal solidification direction. (a) Slow welding speed. (b) Rapid welding speed.

As the weld metal solidifies, impurity elements are rejected into the molten weld pool, eg, sulfur and phosphorus in steel welds (Fig. 7) (8). The final weld metal to solidify, located along the weld centerline at the surface of the weld, has increased levels of these elements, which act to lower the

solidification temperature and, in conjunction with the shrinkage stress resulting from solidification, increase the susceptibility of a weld to solidification cracking at the centerline. This increased risk of cracking is particularly true for fast welding speeds, which promote solidification toward the weld centerline and segregate impurities there.

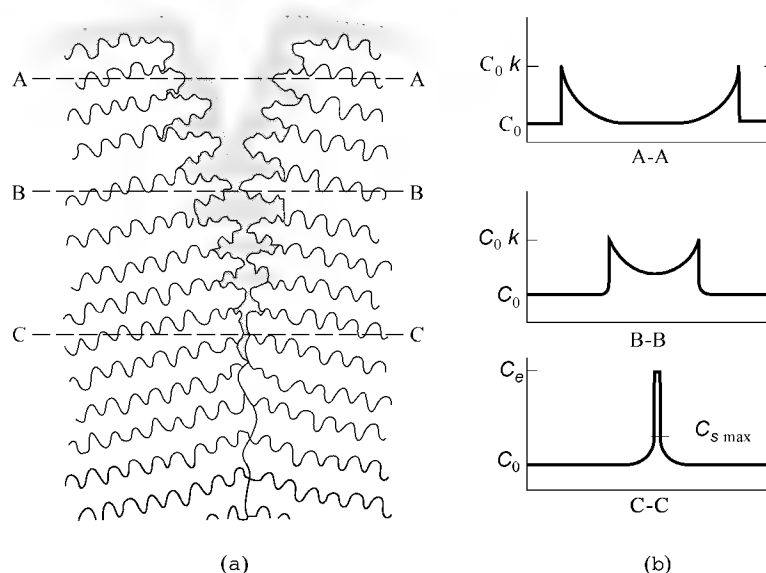


Fig. 7. (a) Impurity elements are rejected into the liquid between the dendritic solidification fronts. (b) Corresponding impurity concentration profiles. C_0 , weld metal composition; k , impurity partitioning coefficient in the liquid; $C_{s \max}$, maximum impurity solid solubility; C_e , eutectic composition at grain boundary.

© NEMJET, The Ohio State University.

The thermal cycle of the welding and solidification process has caused additional complex effects to occur. Differing peak temperatures reached at different locations from the weld significantly affect the microstructure of the base metal in the region surrounding the weld. This region, known as the heat-affected zone, contains metal in various types of transformations, including grain growth and recrystallization. The most pronounced of these occurs at the fusion boundary, where large grain sizes characteristic of the adjacent weld metal are formed. The effects diminish the farther away from the weld one moves until, at some distance, there is no discernible change in the base metal.

Metallurgy. Welding metallurgy deals with the interactions of the base and filler metals and the interactions of these materials with various chemicals injected into the weld via gases, electrode coverings, fluxing and slagging agents, and surface contaminants. For example, a number of gas-metal reactions are possible where the weld is still at high temperature. Oxygen, nitrogen, water vapor, and carbon dioxide are gases that react with ferrous metals to yield products harmful to the metallurgical properties of a weld. The nature of the slag-metal reactions that occur in the molten state strongly depends on the composition of the flux or the electrode coating. Flux chemistry may be altered to control removal of specific weld-metal impurities, such as the addition of manganese or silicon to provide strong deoxidizing action, or additions to enhance slag removal, with all of these additions influencing the final metallurgical characteristics of the weld.

The microstructure of a weld is the overall arrangement of grains, grain boundaries, and phases that exist once solidification occurs. An important metallurgical tool for understanding weld microstructure is the phase diagram, which, for a given alloy composition, relates material phase and temperature. For example, the iron-carbon, Fe-C, phase diagram shows that a steel of 0.25-wt % carbon melts above 1520°C; at a slightly lower temperature it consists of a mixture of molten metal and delta iron; below 1500°C it is transformed to a mixture of liquid metal and austenite; at 1480°C, it solidifies completely to austenite; and below 815°C, a mixture of austenite and ferrite exists, which is transformed to ferrite and cementite, Fe₃C, at 727°C.

However, phase diagrams represent equilibrium conditions that do not prevail in welding. These nonequilibrium conditions result in changes in the temperatures of phase transformations and in microstructures that have solidified before attaining equilibrium. For example, a very rapid cooling of 0.25-wt % carbon steel results in a martensitic microstructure, a material of greater hardness than is achieved under slow cooling. Cooling rates as well as composition are essential to interpreting the microstructures of deposited weld metal and the surrounding heat-affected zone.

Thus, the metallurgy of welds, comprising the weld metal and surrounding heat-affected zone, is influenced not only by the composition of the materials involved, but also by the welding process, the specific procedures for applying the process, and the heat-transfer characteristics (determined by material, mass, and geometry) of the welded joint (9–12).

Material Properties. The properties of materials are ultimately determined by the physics of their microstructure. For engineering applications, however, materials are characterized by various macroscopic physical and mechanical properties. Among the former, the thermal properties of materials, including melting temperature, thermal conductivity, specific heat, and coefficient of thermal expansion, are particularly important in welding. The last property named greatly influences structural distortion that can occur in welding. The electrical conductivity of a material is important in any welding process where base or filler metal is part of the welding electrical circuit.

The response of materials to force is characterized by mechanical properties, eg, elastic modulus, yield stress, tensile strength, ductility, hardness, and impact or fracture strength. Fatigue strength, which is the ability of a material to withstand cyclic loading, is of particular importance to welded structures. Welded joints exhibit greatly diminished fatigue strength, compared to unwelded base metal. The various mechanical properties should be known over a temperature range that covers the expected service temperature. The fracture resistance of most materials, for example, is temperature-sensitive.

Base and Filler Metals

There is hardly a metal that cannot, or has not, been joined by some welding process. From a practical standpoint, however, the range of alloy systems that may be welded is more restricted. The term *weldability* specifies the capacity of a metal, or combination of metals, to be welded under fabrication conditions into a suitable structure that provides satisfactory service. It is not a precisely defined concept, but encompasses a range of conditions, eg, base- and filler-metal combinations, type of process, procedures, surface conditions, and joint geometries of the base metals (12). A number of tests have been developed to measure weldability. These tests generally are intended to determine the susceptibility of welds to cracking.

Base Metals.

Carbon Steels. In addition to iron, these steels contain only carbon and manganese in appreciable quantities as alloying elements. They are mainly used for structural purposes at ordinary temperatures, eg, beams, columns, and storage tanks. Carbon steels are the most easily welded metals. The heat of welding has a metallurgical effect on the base metal that is in accordance with its composition. It can reduce the strength or corrosion resistance of a metal or otherwise change its properties, but this general effect is lower in carbon steels than in other steels. Steels containing up to 0.30 wt % carbon and 1.00 wt % manganese can be readily welded in thicknesses up to ca 50 mm without special techniques. All welding processes discussed above are used to weld carbon steels.

Low Alloy Steels. These alloys are carbon steels to which other elements have been deliberately added to impart a particular property.

Common alloying elements include nickel to improve low temperature mechanical properties; chromium, molybdenum, and vanadium to improve elevated-temperature properties; and silicon to improve properties at ordinary temperatures. Low alloy steels are not used where corrosion is a prime factor and are usually considered separately from stainless steels.

Low alloy steels are readily welded but less easily so than plain carbon steels. Alloying additions increase the hardenability of steel, defined as the ability to form the hard, brittle, nonequilibrium-phase martensite upon fast cooling. Hydrogen, which can come from moisture in the welding consumables or grease on the base metal, can cause cracking in steel when martensite is present. As such, low alloy steels are generally preheated to allow hydrogen diffusion out of the weld and to slow down the weld-metal cooling rate to avoid the formation of martensite. In some cases, it may be necessary to subject the welded joint to heat treatment after welding to relieve residual stress and to temper any martensite that may have formed. All of the welding processes previously mentioned may be used for low alloy steels. The filler metal is usually also low alloy steel.

Stainless Steels. Steels containing a minimum of 12 wt % chromium are referred to as stainless steels. They are broadly categorized, based on their steel microstructures, as austenitic, ferritic, and martensitic; other types are precipitation-hardened and duplex (austenite + ferrite) stainless steels. The austenitic chromium–nickel stainless steels are widely used because of their excellent corrosion resistance. There are about twenty different austenitic chromium–nickel steels, which contain varying amounts of chromium and nickel, eg, 18 wt % chromium and 8 wt % nickel is the most common. Upon welding, these steels can undergo a metallurgical change known as carbide precipitation, which reduces the corrosion resistance of the weld and the heat-affected zone. Stabilized stainless steel grades, to which niobium or titanium has been added, and the extra-low carbon grades may be welded without encountering carbide precipitation. Alternatively, the unstabilized grades may be welded and the weldment subjected to a special heat treatment, which dissolves the precipitated carbides and restores the original corrosion resistance. Solidification cracking has also been associated with these steels, particularly where increased sulfur levels are present. The stainless steels are readily welded by all arc welding processes.

Ferritic and martensitic stainless steels may contain up to 27 wt % chromium. Upon welding, ferritic stainless steels are subject to grain growth in the heat-affected zone, together with a subsequent loss in ductility and toughness. Martensitic stainless steels exhibit low as-welded toughness and ductility, and are subject to hydrogen cracking; a post-weld heat treatment is generally required. These steels may be arc welded using carefully planned procedures, including pre- and post-heat treatment. The filler metal is most often a chromium–nickel steel.

The enhanced strength and corrosion properties of duplex stainless steels depend on maintaining equal amounts of the austenite and ferrite phases. The welding thermal cycle can disrupt this balance; therefore, proper weld-parameter and filler metal selection is essential. Precipitation-hardened stainless steels derive their additional strength from alloy precipitates in an austenitic or martensitic stainless steel matrix. To obtain weld properties near those of the base metal, these steels are heat treated after welding.

Aluminum and Aluminum Alloys. Aluminum alloys are used wherever lightness or atmospheric corrosion resistance are required, or where mildly corrosive fluids are involved. Typical applications for aluminum alloys include railroad tank cars and the skin on aircraft; increasingly, aluminum is being used in the automotive industry. Aluminum alloys are classified as heat-treatable or nonheat-treatable. Strengths of commercially pure and nonheat-treatable alloys are developed by strain hardening and by alloying elements of which magnesium, manganese, and silicon are typical examples. The beneficial effects of strain hardening can be erased by the heat of the welding process; thus, heat inputs are kept low when welding these aluminum alloys. The alloying elements in heat-treatable aluminum alloys are dissolved in the aluminum at high temperature by a process known as solution heat treatment; subsequent heat treatment precipitates these elements as microscopic particles of intermetallic phases, which strengthen the alloy. The welding heat dissolves these particles at or near the weld, thus reducing the strength; properties can be restored by a post-weld heat treatment.

Molten aluminum has an affinity for hydrogen, which results in gas porosity, generally small and well dispersed, as the weld metal solidifies. Elimination of the sources of hydrogen, such as moisture, cleaning fluids, and mill oil, reduces the amounts of porosity. Weld-metal solidification cracking is noted to occur in aluminum alloys that contain significant amounts of copper and magnesium, such as 2024 and 7075; liquation cracking at the heat-affected zone occurs in 6XXX- and 7XXX-series alloys. Aluminum–lithium alloys, the higher elastic modulus and lower density of which have made them of interest to the aerospace industry, have also experienced cracking problems. Gas–metal arc, gas–tungsten arc, and resistance welding are widely used to join aluminum and its alloys.

Magnesium Alloys. In its pure state, magnesium does not have sufficient mechanical strength for structural purposes and must be alloyed with other elements, such as aluminum, zinc, manganese, and zirconium; rare earths and thorium are used for high temperature applications. Magnesium alloys may be divided into room-temperature and elevated-temperature service groups. Certain magnesium alloys are subject to stress corrosion. Weldments subjected to corrosive attack over a period of time may crack adjacent to the weld seams if the residual stresses are not removed. Gas–tungsten arc welding and gas–metal arc welding are recommended for joining magnesium, the former for thinner materials and the latter for thicker materials. Maintaining a protective atmosphere is a critical issue in welding these alloys.

Nickel and Nickel Alloys. This group of metals includes commercially pure nickel and a variety of nickel alloys such as Monel (ca 67% nickel, 30% copper alloys), Inconel (ca 77% nickel, 16% chromium alloys), Incoloy (ca 32% nickel, 21% chromium, 45% iron alloys), and Hastelloy (nickel–molybdenum–iron alloys). Nickel and nickel alloys are used because of their corrosion-resisting properties; the Inconel and Incoloy alloys exhibit good mechanical properties and oxidation resistance at elevated temperatures. Post-weld heat-treatment cracking, also called strain-age cracking, has been associated with heat-treatable nickel alloys. This cracking is particularly pronounced when the combined titanium and aluminum concentration is greater than 5 wt %. Solidification cracking also occurs in these alloys. Nickel alloys can be welded with the shielded-metal arc, submerged-arc, gas–metal arc, or gas–tungsten arc process.

Titanium and Titanium Alloys. Vacuum or inert gas environments are used to weld titanium and its alloys, which have a high affinity for oxygen; commonly used processes are GTAW, PAW, and GMAW, as well as laser and electron-beam welding. Alpha-titanium alloys, which have a hexagonal close-packed atomic structure, have good weldability and are generally welded in the annealed state. Metastable beta-titanium alloys, which have a body-centered-cubic atomic structure, can be welded in the annealed or heat-treated condition. The ductility of welded joints in beta alloys can be low after post-weld aging; therefore these alloys are generally used in the as-welded condition, having good ductility but lower strength. The mechanical properties of alpha + beta titanium alloys can be greatly altered by the welding thermal cycle; weld-metal ductility is generally low. Unalloyed or alpha-titanium filler metal is used to improve ductility in these alloys. T-6Al-4V (6 wt % aluminum, 4 wt % vanadium) has the best weldability among the alpha–beta alloys and accounts for half of industrial titanium usage (3).

Copper and Copper Alloys. The coppers are divided into oxygen-bearing and oxygen-free coppers. Numerous copper alloys are of commercial importance, including those alloys with zinc (brasses), with tin (phosphor bronzes), and with aluminum (aluminum bronzes); all are weldable. In welding copper itself, the copper must be free of oxygen if the joint strength is required to be equal to that of the base metal. Copper alloys can be welded with the shielded-metal arc, gas–metal arc, and gas–tungsten arc process.

Reactive and Refractory Metals. The reactive and refractory metals, originally used in the aerospace industry, are now welded for many applications. The refractory metals, ie, tungsten, molybdenum, tantalum, and niobium, have extremely high melting points, relatively high density, and high thermal conductivity. The reactive metals, ie, zirconium, titanium, and beryllium, have lower melting points and densities, and, except for zirconium, have higher coefficients of thermal expansion. The metals of both groups are difficult to weld. Their high affinity for oxygen and other gases at elevated temperature excludes those processes that utilize fluxes or those in which heated metal is exposed to the atmosphere. Special care must be taken to maintain a protective atmosphere during welding. Small amounts of impurities cause brittleness. The surfaces must be well prepared and very clean to maintain a contamination-free environment during welding and cooling. Beryllium, because of its toxicity, requires special precautions. Gas–tungsten arc welding may be used for all these metals and gas–metal arc welding for some. Electron-beam and laser welding may also be used.

Filler Metals. Filler metals are added to a weld by melting a consumable electrode or a separate wire fed into the weld pool. In the first category, the filler metal is part of the welding electrical circuit and may be in the form of short lengths of covered wire, as in shielded-metal arc welding, or in the form of continuous reels of wire used in semiautomatic, mechanized, and automatic welding processes. Solid wire is used in the gas–metal and submerged-arc welding processes, whereas a hollow, flux-filled wire is used in flux-cored arc welding. More filler metal in the form of iron powder is sometimes added to the electrode coating or flux. In the second category, the filler metal may be in the form of short lengths of bare solid wire, as used in gas welding or manual gas–tungsten arc welding, or in continuous reel form, used in automatic gas–tungsten arc welding.

Filler metals are manufactured in many special forms for welding the commercial alloy systems described herein. The American Welding Society

(AWS) has issued specifications covering the various filler-metal systems and processes (2), eg, AWS A5.28 which applies to low alloy steel filler metals for gas-shielded arc welding. A typical specification covers classification of relevant filler metals, chemical composition, mechanical properties, testing procedures, and matters related to manufacture, eg, packaging, identification, and dimensional tolerances. New specifications are issued occasionally, in addition to ca 30 established specifications. Filler-metal specifications are also issued by the ASME and the Department of Defense (DOD). These specifications are usually similar to the AWS specification, but should be specifically consulted where they apply.

Design

Welded Joints. The weld joint is the geometric arrangement between two pieces of base metal brought together for purposes of welding (13). There are only five recognized weld-joint configurations: corner, butt, tee, lap, and edge joints (Fig. 8). Thus, a butt joint is located between two members in approximately the same plane, whereas a corner joint brings the edges together with an included angle between the planes of the two parts. A tee joints brings the edge of one part onto the planar surface of the second part with a 90° angle between the plates. In a lap joint, one member overlaps another, and an edge joint brings the edges of two parts together without an included angle.

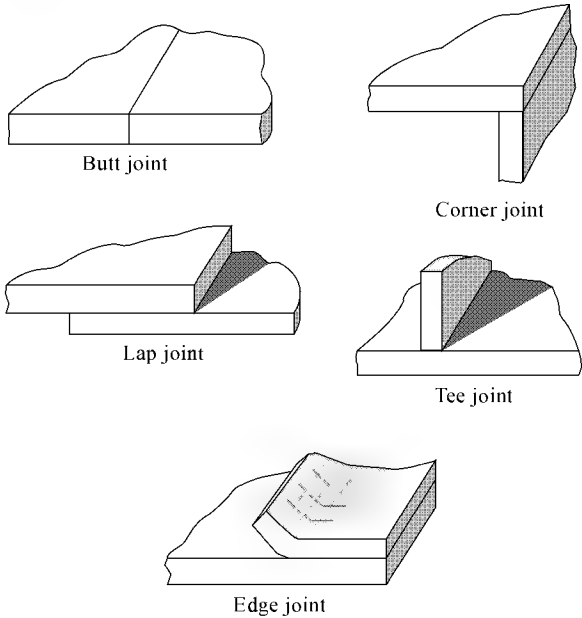


Fig. 8. The five types of welded joints.

Weld Types. A weld may be applied to the various joints in different ways, as shown in (Fig. 9). The fillet weld joins corners and tees. In the plug weld, the two pieces are joined by weld metal deposited in a prepared hole in the overlying piece. For a spot weld, heat is applied to the overlying plate, creating fusion at the interface. Resistance welding is generally used for the spot weld and the seam weld. The groove weld joins base metals in a butt joint. For the backing weld, the root of the original weld is first removed by chipping or gouging. Surface welds are used to build up the surface of parts with special materials or to replace worn material. Flange welds are applied to edge joints, particularly to thin-section materials.

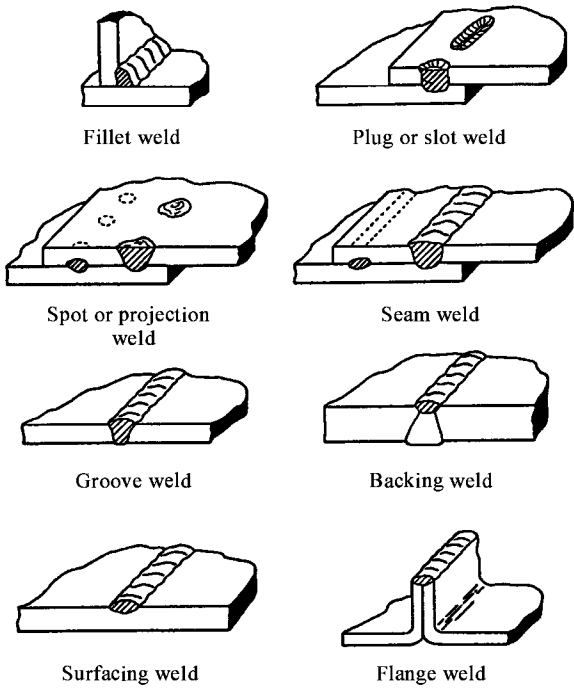


Fig. 9. The eight types of welds.

The edges of the joint to be welded must be prepared. The simplest joint is the square butt joint welded from one or both sides. However, such welds can only be used on thin-gauge material, since it is difficult to penetrate the joint. Various edge penetrations permit fusion across the thickness of the material. Bevel and V grooves are typically prepared by flame cutting, whereas J and U grooves require machining. The selection of single or double grooves is dictated by a variety of practical considerations. Thus, for access to only one side of a surface, a single groove is needed. Double grooves require additional preparation, but their costs are usually offset by a net reduction in weld cross-sectional area, resulting in savings in metal and welding time.

Another design factor is the position in which the weld must be made, which identifies the general orientation of the surface or of the axis of the

weld (2). The positions are referred to as flat, horizontal, vertical, and overhead. In the flat position, welding is performed from the upper side of the joint, and the surface is horizontal. In a horizontal weld, the axis is horizontal, but the surface is inclined at an angle. Most processes are applied to the flat position, which is the most convenient and gives the least number of defects. The most difficult position is the overhead position, and those welds are more prone to defects. Manual or semiautomatic processes require higher skill for welding in these positions. Special fixtures and positioning equipment are often used to permit welding in the flat or horizontal position.

Stress and Distortion. The forces acting on a structure are transmitted through the welded joints; that is, the joint is subjected to simple tension (or compression), bending, shear, or torsional stresses, or to combinations of these stresses owing to combined loading situations. Weldments must be of a proper size, length, and location to withstand the loads imposed during service.

The magnitude and nature of the load are considered in formulating the design. The load may be essentially quasistatic, cyclic, or impact. Many structural failures, for example, have been caused by supposedly innocuous structural details welded in place without any consideration given to their effect on fatigue properties. The service temperatures are also important, since they affect the fracture resistance of a material.

The welding process itself may induce significant stress and distortion, primarily because the molten filler metal is at maximum volume. With cooling and solidification, the material contracts and thereby exerts stresses on the surrounding base metal. Without structural restraint, these stresses cause distortion until they are relieved or minimized. Distortions that occur in simple fillet and butt welds are shown in Figure 10. However, constraint imposed by the surrounding structure may prevent distortion. In this case, significant residual stress may be present, such as tensile stress along the weld axis in a plate. The residual stresses owing to welding may be quite high, approaching or reaching the yield stress of the material. These stresses, in combination with metallurgical considerations, may cause cracking in the weld.

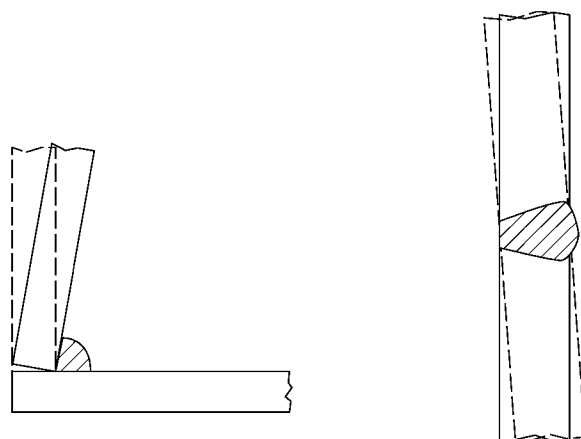


Fig. 10. Weld distortion.

Welding Discontinuities

Welding discontinuities are divided into several categories. Fatigue cracks, creep cavities, and failure resulting from overstress are referred to as service-related. Because steel comprises such a large percentage of welded structures, hydrogen-related discontinuities, which are unique to steel, are listed as a separate category. Fabrication discontinuities occur as a result of the welding process and are described herein. Liquid weld metal can dissolve more gas than solid metal; thus, solidifying weld metal evolves gas bubbles, which may become trapped in the solid metal as porosity. Welding processes that employ a flux are subject to trapped slag inclusions if slag is not completely removed between weld passes. Lack of fusion occurs when weld metal does not properly fuse with either the side of the joint or a previous weld pass. Weld spatter consists of small droplets of electrode material that land beside the weld and adhere to the base metal. Undercut results where a groove, melted at the edge of a weld, is not filled with weld metal. Where the size of a discontinuity exceeds the tolerance limits of a specified welding code, the discontinuity is termed a defect and must be removed, eg, by grinding, and rewelded.

Testing

The integrity of welded structures depends on the integrity of the welds, and much attention is given to testing methods, such as destructive tests, nondestructive tests, and general weld inspection. An objective of many tests is to determine whether welds contain specific defects, such as porosity, slag inclusions, cracks, or lack of fusion (14,15).

Destructive tests destroy the specimen or a portion of the production under examination but provide direct information on properties such as tensile strength, impact strength, ductility, and corrosion resistance. Standard destructive tests include determination of the chemical composition of base or weld metal; corrosion-resistance tests; metallographic tests that employ microscopic examination of polished and etched specimens; and hardness tests across the base metal, heat-affected zone, fusion boundary, and weld metal. Mechanical tests include tensile, impact, and guided bend tests. Destructive tests may be used to evaluate the suitability of welding processes or specific welding procedures for a given application. Simple bend or break tests are used for welder qualification. Test requirements are called out in relevant welding and fabrication codes.

Nondestructive evaluation, also termed nondestructive testing or nondestructive inspection, is extensively used in weld testing (14). Nondestructive tests do no impair the serviceability of the material or component under stress. The most widely used tests for evaluation of welds are liquid penetrant, magnetic particle, ultrasonics, and radiography. Acoustic-emission tests are increasingly used. Nondestructive tests detect and characterize, in terms of size, shape, and location, the various types of weld discontinuities that can occur.

Weld inspection duties of personnel responsible for judging the quality of welding with regard to specifications have been treated (15). Some of these duties involve the visual inspection of welds to determine if they are of the proper size, location, and type and are free of defects. Specifications of materials used must be checked, as must equipment and procedures.

Economic Aspects

More than 1250 companies are involved in the direct manufacture of products associated with welding, such as power sources, electrodes, and fluxes, as well as a wide range of accessories, eg, protective clothing, positioners, and manipulators (16). An extensive distributor network is involved in sales of these welding products. The Standard Industrial Classification (SIC) system of the United States Government lists 14 welding-related industrial groups (17). A search of the *Thomas Register of American Manufacturers* (18) reveals 2455 companies that offer welding services, ranging from arc and resistance welding to the welding of plastics. *Moody's Industrial Manual* indicates that goods using welding in some stage of fabrication account for at least 30% of the gross national product (19).

The U.S. Bureau of Labor Statistics (20) has listed 416,000 persons employed as welders, cutters, and welding machine operators, with 90% employed in the fields of manufacturing, services, construction, and wholesale trades. The same report projects a decline in employment for welders; job prospects remain good, however, as the number of qualified workers entering the market is expected to balance workers leaving the field.

Health and Safety Factors

Welding is carried out in a wide range of industrial environments, such as field construction sites, factory production floors, and small shops. Welders, therefore are subject to the same hazards as all other workers in the metalworking trades. Specific additional hazards of welding include electrical shock, arc radiation, fumes and gases, fires and explosions, compressed gases, cutting and chipping operations, and high noise levels. Many of these hazards also involve workers who share the welding environment. The American National Standards Institute publishes ANSI Standard Z49.1, which covers safe practices for welding and cutting operations (21).

Fumes and gases associated with welding continue to be an area of concern (22–24). Fumes emanate from a number of sources, including electrodes, wires, base metals, coatings and contaminants, ozone produced by the ultraviolet radiation from the welding arc, and gases produced from the heat of the arc. These fumes may lead to a number of health problems, which include acute poisoning under severe cases of ozone and nitrogen oxide concentration, chronic respiratory disease; a condition known as metal-fume fever involving zinc-containing fumes; skin disorders; and disorders of the nervous system, which can be caused by lead or manganese present in welding fumes. A detailed treatment of health and safety in the welding field is found in Reference 25.

BIBLIOGRAPHY

"Welding" in *ECT* 1st ed., Vol. 15, pp. 34–44, by S. A. Greenberg, American Welding Society; in *ECT* 2nd ed., Vol. 22, pp. 241–252, by E. A. Fenton, American Welding Society; in *ECT* 3rd ed., Vol. 24, pp. 502–521, by K. F. Graff, The Ohio State University.

1. *Welding Handbook*, 8th ed., Vol. 1, American Welding Society, Miami, Fla., 1987; see also other volumes in this series.
2. H. B. Cary, *Modern Welding Technology*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1994.
3. *Welding and Brazing*, Vol. 6 of *Metals Handbook*, 10th ed., ASM International, Materials Park, Ohio, 1993.
4. D. Simonson, *The History of Welding*, Monticello Books, Inc., 1969.
5. *Resistance Welding Manual*, 4th ed., Resistance Welder Manufacturers' Association, Philadelphia, Pa., 1989.
6. M. Schwartz, *Metals Joining Manual*, McGraw-Hill Book Company, Inc., New York, 1979.
7. H. Manko, *Solders and Soldering*, 2nd ed., McGraw-Hill Book Company, Inc., New York, 1979.
8. R. J. Bowers and J. C. Lippold, *Introduction to Materials Behavior*, NEMJET, The Ohio State University, Columbus, Ohio, 1996, p. 13.
9. S. Kou, *Welding Metallurgy*, John Wiley & Sons, Inc., New York, 1987.
10. E. Linnert, *Welding Metallurgy*, 4th ed., Vol. 1, American Welding Society, Miami, Fla., 1994.
11. F. Lancaster, *Metallurgy of Welding*, 5th ed., Chapman & Hall, New York, 1993.
12. D. Stout and W. D. Doty, *Weldability of Steels*, 4th ed., Welding Research Council, New York, 1987.
13. W. Blodgett, *Design of Weldments*, The James F. Lincoln Arc Welding Foundation, Cleveland, Ohio, 1963.
14. *Nondestructive Evaluation and Quality Control*, Vol. 17 of *Metals Handbook*, 9th ed., ASM International, Materials Park, Ohio, 1989.
15. *Welding Inspection*, 2nd ed., American Welding Society, Miami, Fla., 1980.
16. *Welding Design Fabricat.*, 17th Annual Welding and Fabricating Buyers Issue (Jan. 1995).
17. *Standard Industrial Classification Manual*, National Technical Information Service, Springfield, Va., 1987.
18. *Thomas Register of American Manufacturers*, Thomas Publishing Company, [http:// thomasregister.com](http://thomasregister.com), 1996.
19. *Moody's Industrial Manual (American and Foreign)*, Moody's Investors Service, Inc., New York, 1981.
20. *Occupational Outlook Handbook*, Bureau of Labor Statistics, [http:// stats.bls.gov oco ocos226.htm](http://stats.bls.gov/oco/ocos226.htm), Apr. 8, 1996.
21. *Safety in Welding and Cutting*, ANSI Z49.1-1988, American Welding Society, Miami, Fla., 1988.
22. Y. Speight and H. C. Campbell, eds., *Fumes and Gases in the Welding Environment*, American Welding Society, Miami, Fla., 1979.
23. *Effects of Welding on Health*, Vols. 1–9, American Welding Society, Miami, Fla., through 1991.
24. *The Facts About Fume*, The Welding Institute, Abington, U.K., 1976.
25. N. Balchin and H. Castner, *Health and Safety in Welding and Allied Processes*, 4th ed., McGraw-Hill Book Co., Inc., New York, 1993.

Randy J. Bowers
Karl F. Graff
Edison Welding Institute

WETTING AGENTS.

See SURFACTANTS; DETERGENCY.

WHEAT AND OTHER CEREAL GRAINS

Origins and History

The word cereal derives from the name of the Roman goddess, Ceres, in whose honor a spring festival, The Cerealis, was celebrated. This indicates the antiquity of these foods and the reverence with which they were esteemed. The latter probably reflected the important role cereals played in the diets of ancient peoples. Cereals are still an important dietary ingredient. As the world population continues to grow, cereals will become an increasing fraction of the diets for more and more people. That is recognized by the use of an ear of wheat as the symbol of the FAO. Below the wheat is the Latin inscription, *Fiat panis* (Let there be bread) (1). Exactly where and when cereal grains were first cultivated has not been established. Most reports suggest that, for wheat, cultivation first occurred somewhere in the "fertile crescent" which extends from Egypt through Syria to and including the Tigris–Euphrates valley (2). It was in that region that the likely progenitors of wheat, emmer and einkorn, were found. These wild grains still grow in the area extending eastward from Asia Minor to Iran and Afghanistan.

The origins of the other cereals are less certain. Rice may have originated in Africa or Asia but probably was first cultivated somewhere between the southern People's Republic of China and southern Vietnam (3). Corn or maize originated in the Western Hemisphere. The Spanish conquistadors came in contact with corn during their early explorations of Central and South America (4). Corn originated perhaps in the lowlands of South America; others place its origin in Mexico or Guatemala (5). Wherever it was developed, by the time Columbus made his first voyage of exploration, corn as a food crop had spread over much of the Americas as well as to the West Indies (5). By 1492, the Indians had developed corn culture to such a high state that it is thought to have then ranked highest among cereals in efficiency of food production (6). The origin of rye is more difficult to ascertain. There appears to be no reference to it prior to the Christian era. It is still a principal ingredient in the diet of some peoples in northern Europe as it was in the U.K. until the late eighteenth century (7). Since barley grows wild in the Syria–Palestine area, it is likely to have originated there (7). Millet may have originated in the Sudan where pearl millet is still widely cultivated (3).

Botanically, cereals are simple, one-seed fruits. They include wheat, rice, corn (also called maize in some parts of the world), rye, barley, oats, sorghum, and millet. All cereals contain large amounts of starch and little fat; the fat is associated primarily with the germ and scutellum (single cotyledon or the first leaf). In most cases, the lipids in cereals contain a high concentration of unsaturated fatty acids, which are protected from oxidation by the presence of tocopherols. Actually, the oil from wheat germ was the starting material for the isolation of α -tocopherol (see VITAMINS, VITAMINE). As long as the antioxidants are in proximity to the lipids, oxidation of the lipids and the accompanying rancidity is minimized. This is one reason why wheat is stored in the kernel stage until it is milled. Generally, wheat is milled only when there is an order for flour. However, even in the intact kernel, the lipids are subject to hydrolytic rancidity (8), which together with oxidative rancidity can be estimated by determining the free fatty acid content. That is done by determining the amount in milligrams of potassium hydroxide required to neutralize the free fatty acids from 100 g of moisture-free grain (9). This is called the fat acidity value. One of the main purposes of milling, in addition to producing an acceptable flour, is to remove the wheat germ in as intact a state as possible. Thereby, the concentration of the lipids is reduced, the development of oxidative rancidity decreased, and the shelf life of the flour extended.

Cereals were among the earliest plants cultivated. They were related to some of the wild grasses indigenous to those parts of the world where civilizations had their origins. All of them are ideally suited for use as food under both primitive, and advanced conditions. For one thing, they can be stored for long periods under many conditions, thus providing a reserve against food shortages. This was done in ancient Egypt when Joseph advised the Pharaoh to store grain during the seven "plenteous" years to be used in the following seven "lean" years. In addition, cereals are nutritious foods that can be used in many ways, thus facilitating their incorporation in the diet at high levels over long periods of time. Even today, there are some areas, such as rural Iran, where wheat products, especially in the form of the flat breads indigenous to that region, provide as much as 70–90% of the daily caloric intake. Most of the cereals also respond well to primitive methods of agriculture with good yields and, with advanced technology and improved varieties, large yields per worker can be secured. Much more food is secured from fields planted in grains than can be obtained from cattle or other animals on the same land. The importance of this became evident during World War I. In the early stages of that conflagration, the German High Command made a conscious decision to continue the prewar level of meat, milk, and egg production to provide adequate nutrition for the men in the armed services and for the civilians who would be called on for arduous work in connection with the war. Had they decided instead on conversion of meadows and pastures to wheat fields and the direct consumption of the grain by human beings, the yield of food for the German people would have been far greater than it was when the land was given over to raising cattle, pigs, sheep, and poultry. This led a group of Scottish physicians to suggest that this mistake probably did more than any army general to lose the war (10). Finally, cereals grow in a wide variety of climatic and soil conditions, and they successfully compete with weeds for the limited amounts of nutrients and water where these plant factors are in short supply. This is an important reason why cereals have played, and continue to play, such an important role in the development of the human race.

Nutritional Value of Cereals

Deficiency Diseases. Not only did cereals make an important contribution to improving the general status of humankind, but they also were important dietary components of some groups of people who showed certain nutritional deficiencies. This observation led to the discovery of some of the vitamins. These deficiency diseases have been most prominently associated with use of rice, corn, and wheat.

Beriberi, Thiamine Deficiency. The recognition of vitamins and their importance to the health of human beings came about when Eijkman, a Dutch pathologist, was sent to Java in an attempt to cure an epidemic of beriberi that had appeared in one of the hospitals. Eijkman kept a flock of chickens on the hospital grounds to assist in discovering the disease agent he assumed was involved in the etiology of beriberi. These chickens were fed the scraps from the plates of the hospital patients—primarily polished rice, the common food in that part of the world (11).

Although Eijkman recognized the condition in his polyneuritic chickens as the analogue of beriberi, he misinterpreted its cause. He suggested on the basis of bacteriological concepts then dominating the medical field that beriberi resulted from the ingestion of a toxic substance associated with the starchy part of the rice. According to that theory, rice polishings contained an antitoxin which neutralized the toxin present in the polished rice. It was Eijkman's successor who established that polished rice lacked some substance essential for normal physiological functioning and that this substance occurred in rice polishings, beans, meat, and other foods (12). The critical ingredient is now known to be thiamine (vitamin B₁). The absence from the diet of similar substances is responsible for the development of scurvy, pellagra, rickets, etc (13).

Pellagra, Niacin Deficiency. It was 220 years after the first description of pellagra that nicotinic acid was discovered to be the cure for black tongue in dogs (14), a condition suggested by a veterinarian in North Carolina to be similar to human pellagra (15).

The contrast between a high incidence of pellagra among the inhabitants of the southeastern United States and its absence among the corn-eating people south of the Rio Grande was a puzzle for a number of years. It finally became evident that the lime water in which the corn kernels were steeped prior to being made into tortillas liberated nicotinic acid from the bound form (16). Untreated kernels of corn contain nicotinic acid as a complex from which it cannot be made available by human or animal digestive processes. However, a weak alkali, such as lime water, releases nicotinic acid from its complex and makes it available for absorption. Furthermore, the usual diet of the corn eaters south of the Rio Grande includes a large complement of beans. Beans are a rich source of tryptophan which can be converted to nicotinic acid by enzymes in the human body (see VITAMINS, NIACINE, NICOTINAMIDE, AND NICOTINIC ACID).

Zinc Deficiency. A nutritional problem associated with consumption of large amounts of whole wheat products is the unavailability of dietary zinc, first observed in a patient with immature development and dwarfism in southern Iran (17). The patient showed marked improvement when placed on a well-balanced, nutritious diet for a year. In 1962, similar patients were observed in Egyptian villages. A deficiency of zinc was identified as the primary reason for the development of this condition. This deficiency of zinc results from the binding of that metal by the phytates present in whole wheat (18). Even when an excess of zinc is present in the diet, if conditions are right, that zinc may be complexed with the phytate and thus rendered unavailable to the body (see MINERAL NUTRIENTS). Thus, a zinc deficiency may develop when its dietary level appears adequate but the diet contains large amounts of a food, such as whole wheat, that is high in phytates.

Calcium Absorption. Phytates in cereal grains have also been reported to interfere with the absorption of calcium. However, a long-term study indicated a retention of calcium in subjects that consume large amounts of bread made with high extraction of flour (19).

Lysine Concentration. Many nutritionists are concerned about the low concentration of lysine in wheat proteins (see AMINO ACIDS). Initially, the insufficiency of lysine in wheat was noted because of the failure of weanling rats to grow when fed a ration in which white wheat flour provided most or all of the protein in the diet. The rats grew only when the white flour ration was supplemented with lysine or a protein source containing a relatively high concentration of that amino acid. Animal proteins were found to be good sources of lysine. That led nutritionists to advocate the use of wheat products in conjunction with animal proteins in human diets. The idea that human beings require not only cereals but also animal foods in their diet provided the theoretical basis for the development of American agriculture and much of the nutritional advice given the public.

However, the validity of rat studies for human nutrition has been called into serious question because of the vast difference in rates of growth of rats and humans (20). The daily weight gain of the infant rat is about 3% of its weight; the comparable rate for the human infant is 0.03%. The 100-fold difference in relative weight gains makes it impossible to evaluate food intended for human beings on the basis of its ability to provide good growth in rats (21).

Health Advantages of High Cereal Consumption

Blood Urea Levels and Kidney Function. One of the most intriguing observations made during a seven-week "bread" study was the reduction in blood urea level of the subjects (22). In normal individuals, the blood urea level was believed to be directly proportional to the amount of protein consumed (23). The normal blood urea concentration of individuals consuming 70 g of protein per day was projected as 32 mg 100 mL, and that was the level observed in the subjects at the end of the three-week control period. However, sometime after the "bread" diet was started, the blood urea level dropped to one half of the control value even though there had been no change in the level of dietary protein (22). Although there had been no change in the intake level of protein, the study did involve a complete elimination of animal protein during the "bread" diet period.

A marked reduction in the blood urea level of normal individuals implies an improvement in kidney function (see UREA; DIURETIC AGENTS). Such was observed in a subsequent study (24). Urea and creatinine clearances (clinical measures of kidney function) of six normal college men showed a 25% and 10% improvement, respectively, when they consumed a diet high in bread for seven weeks. This improvement occurred even though the tests at the end of the three-week control period were normal. Throughout the study, the protein level in the diet was maintained constant at 70 g d. During the control period, the protein came from a mixture of plant and animal sources but during the "bread" period, ca 90% of the dietary protein came from wheat and animal protein was completely excluded. The improvement in kidney function associated with the consumption of a diet high in wheat products suggests a possible therapeutic value for wheat. That might be especially true for those patients who are starting to manifest aberrations in kidney function which ultimately would require renal dialysis or a kidney transplant.

Prevention of Osteoporosis. Another health benefit associated with the consumption of a diet high in cereals is the maintenance of bone-mineral content, especially among older women. After the menopause, women are subject to demineralization of bones, especially in the pelvis and spinal column. This condition, called osteoporosis, leaves them vulnerable to hip fractures. Some degree of deterioration in total life adjustment is a nearly inevitable occurrence in the five sixths who survive a hip fracture (25). Osteoporosis in women has been associated with a reduced estrogen production after the menopause. That plus the fact that animal studies indicated an involvement of estrogen in depositing calcium in the bones led to use of this hormone in treatment of osteoporotic women (26). Reports suggest, however, a connection between estrogen therapy and endometrial cancer (27).

Since dietary cereals are low in sulfur-containing amino acids, they produce an alkaline urine which favors the retention of bone minerals. In post-menopausal women, there appears to be some interaction between the diet and the effect produced by estrogens on bone mineral content (28).

Prevention of Obesity. Cereal products may have an important role in obesity control because of the low concentration of fat in cereals and the fact that the starch in wheat flour acts, in the human body, somewhat like dietary fiber (29). Both the restriction in dietary fat and the presence of dietary fiber aid the individual in reducing caloric intake with a minimum of effort, and decrease the hunger reaction that plagues most people attempting to lose weight. Many people believe bread and cereal products and fattening, although bread frequently serves as the innocuous vehicle for the conveyance to the mouth of various fattening foods.

When the intake of fatty foods is curtailed, a large amount of bread in the diet is likely to produce a weight loss. For example, eight college men, eating 12 slices of commercial white bread per day, lost an average of 6.3 kg over an eight-week period (30). All of these men were from 10–15% overweight. Despite the continued loss of weight over the eight-week period, the few men in the group who experienced hunger reported it to be mild (30).

Dental Caries and Isomerase. Susceptibility to dental caries is a genetically determined trait. In susceptible individuals, the incidence of dental caries is directly related to the amount of sucrose in the diet (31–33); however, there are a number of factors that may modify the incidence. Although sucrose in the diet produces a high incidence of dental caries, its component monosaccharides used at an equivalent level in the diet result in a much lower incidence (see SUGAR). It appears that the oxygen bridge linking glucose and fructose to form sucrose is needed as a source of energy by the bacteria (probably *Streptococcus mutans*) responsible for the development of the carious lesion.

The development of high fructose corn syrup (HFCS) may provide another health benefit attributable to cereal grains (see SYRUPS). These syrups are being used to an ever-increasing extent by the food industry. Shortly after HFCS came on the market, a dramatic increase in the price of sucrose acted as a spur to the production of the high fructose syrups.

The high fructose syrups are produced from corn starch after the latter has been hydrolyzed to glucose by either acid or α -amylase. Treatment of the resulting fluid with an isomerase converts ca 42% of the glucose to fructose. The remaining solids consist almost exclusively of glucose and other saccharides (34). After clarification, this product is concentrated to a base of 71% solids, in which state it is bulk-distributed. The enzymes involved in the isomerization are of bacterial origin, most of them belonging to the *Streptomyces* group. The enzyme is immobilized by fixation on cellulose (see ENZYME APPLICATIONS).

A solution is available with a 90% fructose content. Because of the sweetness of fructose, this syrup imparts the same sweetness to a product as a larger amount of sucrose (see SWEETENERS). Since on a weight basis the monosaccharides have the same number of food calories as sucrose, this gives a means of reducing the caloric content of the food. Reduced caloric foods, such as jellies and preserves, salad dressing, and table syrup, are now being manufactured with 90% fructose corn syrup and are recognized as having quality improvements compared to the use of other sweeteners (34).

Health Problems Associated with the Consumption of Cereals

Celiac Disease. A disturbance of the lower gastrointestinal tract, celiac disease is a chronic disease characterized by loss of appetite and weight, depression and irritability, and diarrhea frequently followed by constipation (35). One of the more disturbing features of celiac disease is the large, frothy, foul-smelling stools. The disease may develop in childhood or later in life. Frequently, the patients who develop the disease in adulthood report having had some of the symptoms during childhood.

This disturbance was recognized shortly after World War II as being related to the ingestion of wheat. A group of physicians in the Netherlands was impressed by the fact that during the war they saw many cases of celiac disease. During that time, wheat was the primary staple of the diet. However, at the end of the war, other foods again became available and the number of children who developed celiac disease decreased. One Dutch group of investigators had a seven-year-old female patient who displayed extreme fluctuations in her symptoms. These changes were shown to be associated with the presence or absence of bread in her diet. Using that patient as the test subject, the group soon learned that her symptoms worsened shortly after she consumed foods containing wheat gluten (36). Although considerable work has been done attempting to identify the mechanism whereby gluten causes the disturbance characterizing celiac disease, little more can be said beyond that this appears to be an allergic reaction. For victims of celiac, the primary therapy involves the elimination of any dietary product that contains gluten. Since that protein occurs not only in wheat but also in rye and barley, the complete elimination of gluten from the diet becomes an onerous task.

Disturbances Associated with Flour-Aging Agents. Bread made with flour freshly milled from wheat has a different texture and appearance than that associated with bread. Only after the flour has been aged by storage for a few months does the gluten develop the toughness that permits the bread to retain the shape it had when it was put in the oven. The aging of the flour likely involves a mild reaction of oxygen and some part of the gluten molecule. Additional evidence for this is the destruction of the yellowish pigment present in the flour prior to its storage. To hasten the reactions that occur during storage, chlorine gas and oxides of nitrogen were tried. Although they improved the freshly milled flour, they were not completely satisfactory. In 1921, nitrogen trichloride was introduced as a flour agenizing substance (37).

Although no untoward effects were observed that could be attributed to the long-term ingestion of agenized flour by people in the United States and the U.K., and despite the absence of any symptoms in human subjects given large amounts of flour heavily treated with nitrogen trichloride prior to incorporation into foods, the use of nitrogen trichloride by the milling industry in the United States was prohibited after August 1, 1949 (38). That order recommended the use of chlorine dioxide for agenizing flour. Like nitrogen trichloride, chlorine dioxide is unstable and must be passed into the flour immediately after being generated. The chlorine dioxide is formed by passing chlorine mixed with air into a solution of sodium chlorite (39) (see CHLORINE OXYGEN ACIDS AND SALTS).

Ergot. Ergotism is only indirectly related to health hazards associated with the consumption of cereals because the alkaloid which is responsible for ergotism is produced by a fungus (*Claviceps purpurea*) that grows primarily on rye. That the fungus may grow on other cereals than rye is implied in a report of this disturbance among the members of a British family who consumed no rye bread (40) (see ALKALOIDS).

The disease takes two different forms depending apparently on whether ergotoxin, the alkaloid in ergot, attacks predominantly the nervous or the circulatory system. The former condition is characterized by severe convulsive seizures; the latter produces an intense burning and itching of the skin called St. Anthony's Fire.

Abortion is one of the prominent characteristics of ergot poisoning among both women and cattle. This is due to the action of ergotoxin in contracting the placental muscles. For that reason, small, regulated doses of ergotoxin are used to control bleeding following childbirth (see also HORMONES, POSTERIOR PITUITARY HORMONES).

Colonic Cancer. Another area in which cereals may play an important role in maintaining the health of human beings is as a source of dietary fiber (qv). During the past decade, evidence has been collected that certain diseases, to which people in the more-developed countries are prone, owe their origin to the almost complete removal on the fibrous parts of cereals consumed in the diet. There is an almost complete absence of a variety of tumors in the lower gastrointestinal tract among the natives of African villages, whereas the incidence of these colonic tumors among those of European extraction living in the area is as high as in northern Europe (41). The primary, visible difference between these two groups of people is the nature of their diets. The native Africans consume a diet composed largely of unrefined foods and consequently high in fiber. On the other hand, the neighboring people of European origin rely on foods similar to those available in their countries of origin. The foods eaten by the latter group are either devoid of, or very low in, dietary fiber.

Prior to this work, dietary fiber, of which cellulose is one of the more important constituents, was considered important primarily as a means of preventing or overcoming constipation. Otherwise, dietary fiber was considered to be a metabolically inert substance. A large variety of diseases such as appendicitis, hiatus hernia, gallstones, ischemic heart disease, diabetes, obesity, dental caries, and duodenal ulcers are now suspected to be associated with the consumption of a highly refined diet (42).

Diabetes. Fiber has also been shown to be an aid in treating diabetes. Diabetic patients who required 20 or less units of insulin per day could do without their insulin injections provided they increased the amount of cereal bran in their diets (43). These patients had adult onset diabetes, a milder form of the disease than that which appears in children. Nevertheless, the fact that a high intake of bran relieves these patients of the need for daily insulin injections is a welcome form of therapy. Practically all evidence indicates that dietary fiber decreases the maximum blood glucose level following a dose of glucose as in the typical glucose tolerance test. The lower level of blood glucose presumably occurs because its absorption from the intestinal tract is slowed. The mechanism whereby dietary fiber decreases the rate of absorption of glucose from the intestinal tract is not completely understood. One possibility is that the large molecules of dietary fiber physically impede the movement of the glucose molecules to the absorptive surface of the intestine. Another theory suggests that dietary fibers increase the viscosity of the intestinal contents and thereby slow the movement of glucose molecules (44).

Cardiovascular Diseases. For sometime it was believed that dietary fiber protected the individual from various heart disturbances, especially ischemic heart disease. That presumably occurred as a result of the action of dietary fiber in lowering serum cholesterol levels. This effect results, according to reports, from the increased rate of passage of food through the intestinal tract and consequent incomplete absorption of dietary cholesterol and some bile acids. Since some of the bile acids are secreted into the lower intestinal tract to be reabsorbed at a lower site and since cholesterol is the precursor of bile acids, this would lower blood cholesterol levels. However, some investigators now suggest that it may be other differences in dietary constituents or lifestyle that are responsible for the low blood cholesterol levels among primitive people whose diet is high in fiber (45). Inclusion of large amounts of high fiber foods in the diet may decrease the intake of cholesterol-containing foods and those high in saturated fats. The intake of these two dietary components is probably closely related to the level and nature of blood cholesterol (45).

The implications of the role that dietary fiber may play in the maintenance of human health should increase the demand for cereal products. It may well have an influence on the milling industry, especially if the demand for cereal products that are less highly milled becomes a reality.

Wheat: Production, Trade, and Uses

Wheat is cultivated in most countries on all continents. World wheat production, consumption, and net exports (imports) are summarized in Table 1.

Table 1. Wheat: World Production, Consumption, and Net Exports, 1985/1986 (million metric tons)^{a, b}

Country	Production	Consumption	Net Exports
	<i>Principal Exporters</i>		
United States	64.68	30.21	32.53
Canada	23.50	5.50	17.50
Australia	17.00	3.10	15.20
EC-10	70.07	53.07	14.30
Argentina	11.50	4.85	6.70
Turkey	13.00	13.70	0.15
	<i>Principal Importers</i>		
USSR (former)	83.00	100.00	19.00

China	87.00	94.00	−7.00
Eastern Europe	38.98	38.97	−0.21
other Western Europe	9.57	9.68	0.29
Brazil	2.20	6.40	4.40
Mexico	4.40	4.50	0.30
Other Latin America	0.05	2.72	2.69
Japan	0.79	6.35	5.50
India	45.00	43.00	1.40
South Korea	0.02	2.37	2.30
Indonesia	0	1.37	1.15
other Asia	17.56	25.21	8.14
Egypt	1.85	8.45	6.70
Morocco	1.83	4.14	2.30
other northern Africa–Mideast	11.83	25.47	14.23
other Africa	3.58	9.48	5.85
Residual	2.43	10.34	3.92
<i>Total world</i>	<i>509.84</i>	<i>502.88</i>	

^a *Source:* Grain and Feed Division, USDA.

^b Trade on July–June years.

The top five wheat producing countries are the former Soviet Union, the People's Republic of China, the United States, India, and Canada. Of these five countries, only the United States and Canada grow more wheat than they use and export to other countries. The other three nations are large wheat importers. Some of the top wheat customers of the United States have been China, India, the former Soviet Union, Japan, and Brazil.

A nation of one billion people, China is traditionally regarded as a rice-eating nation. But China grows almost as much wheat as the United States and buys and uses more wheat than any other country in the world. Each person in China on the average consumes 180 lb of wheat every year, mostly in the form of noodles. The average American eats only about 116 lb of wheat flour per year in all types of wheat-based products. Some nations have much higher per capita consumption, up to 300 lb of wheat per year per person (46–48).

Wheat is grown in most of the 50 states of the United States. The kind of wheat grown and quantity vary widely from one region to another. In total over 200 varieties are grown annually. Winter wheats are planted in the fall. After the grass-like seedlings emerge from the ground, they lie dormant during the winter. They come up again in the spring, ripen, and are harvested in early summer. Spring wheats are planted in the spring and harvested in late summer. Spring wheats grow best in the northern areas of the United States where the summers are not too hot for the young plants. Winter wheats grow best in areas of the country where the winters are not too harsh for the young plants.

The many varieties of winter and spring wheat are grouped into five official classes. The class a variety fits into is determined by the hardness, the color of its kernels, and its planting time. Each class of wheat has its own relatively uniform characteristics, including those related to milling, baking, or other food uses (46–48). Protein range and flour use of principal wheat classes are listed in Figure 1 (49).

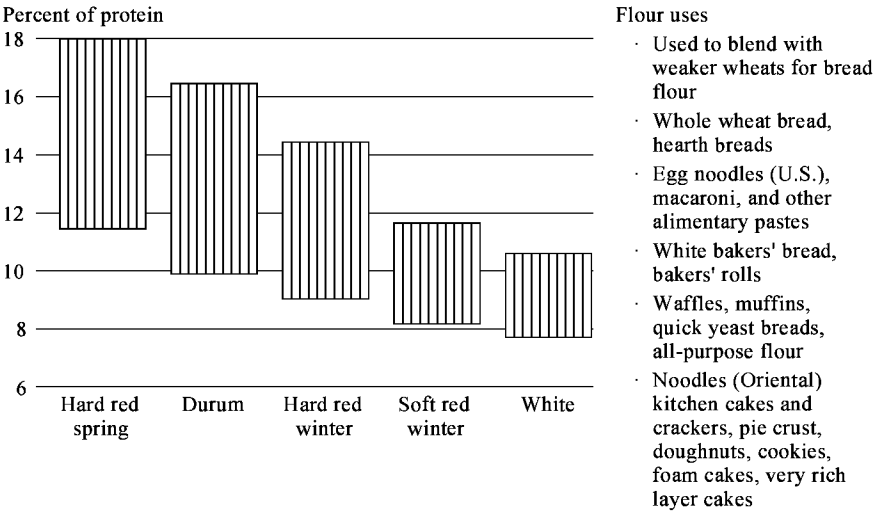


Fig. 1. Protein range and flour uses of principal wheat classes. Flour uses are listed according to the approximate level of protein required for specified wheat products. Durum is not traded on the basis of protein content. From Ref. 49.

Hard red winter (HRW) is an important bread wheat that accounts for more than 40% of the United States' wheat crop and wheat exports. This fall-seeded wheat is produced in the Great Plains, which extend from the Mississippi River west to the Rocky Mountains, and from the Dakotas and Montana south to Texas. Significant quantities are also produced in California. HRW has moderately high protein content, usually averaging 11–12%, and good milling and baking characteristics.

Hard red spring (HRS), another important bread wheat, has the highest protein content, usually 13–14%, in addition to good milling and baking characteristics. This spring-seeded wheat is primarily grown in the north central United States—North Dakota, South Dakota, Minnesota, and Montana. HRS constitutes about 15% of U.S. wheat exports. Subclasses based on the dark, hard, and vitreous (DHV) content, include Dark Northern Spring, Northern Spring, and Red Spring.

White wheat (WW) is a preferred wheat for noodles, flat breads, and bakery products other than loaf bread. WW, which includes both fall- and spring-seeded varieties, is grown mainly in the Pacific Northwest. This low protein wheat, usually about 10%, comprises about 15% of U.S. wheat exports, destined primarily for East Asia and the Middle East. Subclasses include hard white, soft white, western white, and white club.

Soft red winter (SRW), which is grown in the eastern third of the United States, is a high yielding wheat, but relatively low in protein, usually about 10%. SRW best provides flour for cakes, pastries, quick breads, crackers, and snack foods. This fall-seeded wheat constitutes about one-quarter of U.S. wheat exports.

Durum, the hardest of all U.S. wheats, provides semolina for spaghetti, macaroni, and other pasta products. This spring-seeded wheat is grown primarily in the same northern areas as hard red spring, but small winter sown quantities are also grown in Arizona and California. Durum represents about 5% percent of total U.S. wheat exports. Subclasses are hard amber durum, amber durum, and durum.

Grades and grade requirements for all U.S. wheat classes are given in Table 2. Values of the total domestic consumption of wheat products as foods in the United States are listed in Ref. 50.

Table 2. Official U.S. Standards for Grain (USDA 1985): Wheat

	Minimum limits of		Maximum limits of						
	Test weight per bushel, lb		Heat-damaged kernels, ^o %	Damage to kernels, (total), ^o % ^c	Foreign material, ^o %	Shrunken and broken kernels, ^o %	Defects (Total), ^o % ^d	Wheat of other classes ^a	
	Hard Red Spring Wheat or White Club wheat ^b	All other classes and subclasses						Contrasting classes, ^o %	Wheat of other classes, ^o % ^e
U.S. No. 1	58.0	60.0	0.2	2.0	0.5	3.0	3.0	1.0	3.0
U.S. No. 2	57.0	58.0	0.2	4.0	1.0	5.0	5.0	2.0	5.0
U.S. No. 3	55.0	56.0	0.5	7.0	2.0	8.0	8.0	3.0	10.0
U.S. No. 4	53.0	54.0	1.0	10.0	3.0	12.0	12.0	10.0	10.0
U.S. No. 5	50.0	51.0	3.0	15.0	5.0	20.0	20.0	10.0	10.0
U.S. Sample grade ^f									

^a Unclassed wheat of any grade may contain not more than 10⁰ % of wheat of other classes.

^b These requirements also apply when Hard Red Spring wheat or White Club wheat predominate in a sample of mixed wheat.

^c Includes heat-damaged kernels.

^d Defects (total) include damaged kernels (total), foreign material, and shrunken and broken kernels. The sum of these three factors may not exceed the limit for defects.

^e Includes contrasting classes.

^f U.S. Sample grade shall be wheat that (a) does not meet the requirements for U.S. grades 1, 2, 3, 4, or 5; or (b) contains 8 or more stones, 2 or more pieces of glass, 3 or more crotalaria seeds (*Crotalaria* spp.), 2 or more castor beans (*Ricinus communis*), 4 or more particles of an unknown foreign substance(s) or a commonly recognized harmful or toxic substance(s), or 2 or more rodent pellets, bird droppings, or equivalent quantity of other animal filth per 1,000 g of wheat; or (c) has a musty, sour, or commercially objectionable foreign odor (except smut or garlic odor); or (d) is heating or otherwise of distinctly low quality.

Milling of Wheat

Conventional, Modern Milling. Before wheat can be used in the production of most foods, it must undergo several mechanical and chemical changes. The first change involves milling of wheat into flour (46,51). The steps involved in wheat-to-flour production are wheat selection, blending, cleaning, conditioning, milling, and maturing. The endosperm, which forms about 83^o % of the kernel, is the source of white flour and contains 70–75^o % of the kernel's protein. The bran, forming about 14^o % of the kernel, is included in whole wheat flour, but is more often removed and used in animal or poultry feed. Because the cellulosic material of the bran cannot be digested and tends to accelerate the passage of food through the human digestive tract, the total nutritive contribution of whole wheat flour is less than that found in enriched white flour products. The germ, forming about 3^o % of the kernel, is the embryo or sprouting tissue of the seed. It is usually separated out because it contains oil, which limits the keeping quality of flours. Although the germ is available as human food, it is usually added to animal or poultry feed.

The mill flow diagram begins with a separator, where the wheat first passes through a vibrating screen that removes straw and other coarse materials, and then over a second screen through which drop small foreign materials like seeds. An aspirator lifts off lighter impurities in the wheat. After the aspirator, wheat moves into a disk separator, consisting of disks revolving on a horizontal axis. The disk surfaces are indented to catch individual grains of wheat but reject larger or smaller material. The blades push the wheat from one end of the machine to the other. The revolving disks discharge the wheat into a hopper or into the continuing stream. The wheat then moves into a scourer—a machine in which beaters attached to a central shaft throw the wheat against a surrounding drum. Scourers may be either horizontal or upright, with or without brushes, and adjusted for mild, medium, or hard scouring. Air currents carry off the dust and loosened particles of bran coating.

The stream of wheat next passes over a magnetic separator that pulls out iron and steel particles. A washer–stoner may be the next piece of equipment. High speed rotors spin the wheat in a water bath. Excess water is thrown out by centrifugal force. Stones drop to the bottom and are removed. Lighter materials float off, leaving only the clean wheat. In the production lines of some mills, a dry stoner is also used. The wheat passes over an inclined, vibrating table that pushes stones and heavy material away from the lighter wheat, discharged separately.

The clean wheat is then tempered before the start of grinding; in the process moisture is added. Tempering aids the separation of the bran from the endosperm and helps provide a constant, controlled amount of moisture and temperature throughout the milling process. The percentage of moisture, length of conditioning, and temperature are the three important factors in tempering, with different requirements in soft, medium, and hard wheats (46). Some water for tempering may be added, dampening the wheat; the wheat may be washed until a certain amount of moisture is absorbed; or the wheat may be subjected to steam under low pressure. The dampened wheat is held in a bin for a prescribed period, usually 8–24 h, depending on the type of wheat. The outer layers of the wheat berry tend to be brittle, and tempering toughens the bran coat for better separation of the endosperm. Within the kernel, tempering also mellows the endosperm so that the floury particles break more freely in milling. When the moisture content is properly dispersed in the wheat for efficient milling, up to approximately 16^o %, the grain is passed through an Entoleter scourer–aspirator as a final step in cleaning. Disks revolving at high speed in the scourer–aspirator throw the wheat against fingerlike pins. The impact cracks unsound kernels (including insect-damaged), which are rejected. The sound wheat flows to a grinding bin or hopper from which it is fed in a continuous metered stream into the mill itself (46).

The first-break rolls of a mill are corrugated rather than smooth like the reduction rolls that reduce the particles of endosperm further along in the process. The rollers are paired and rotate inward against each other and at different speeds. The clearance between rollers and the pressure as well as the speed of each separate roller, can be adjusted. At each breaking step, the miller selects the milling surface and the corrugations; the speed of and interrelation between the rollers depend on the type and condition of the wheat.

The next important step introduces the broken-ground particles of wheat and bran into a sifter where they are shaken through a series of bolting cloths or screens to separate the larger from the smaller particles. The fractions, and particles of endosperm graded by size are carried to purifiers. In a purifier, a controlled flow of air lifts off bran particles while bolting cloth at the same time separates and grades coarser fractions by size and quality. Four or five additional break rolls, with successively finer corrugations and each followed by a sifter, are usually used to rework the coarse stocks from the sifters and reduce the wheat particles to granular middlings as free from bran as possible. Germ particles, being somewhat plastic, are flattened by passage through the smooth reduction rolls and can be separated. The reduction rolls reduce the purified, granular middlings or farina to flour.

The process is repeated until the maximum amount of flour is separated, consisting of at least 72^o % of the wheat. This flour is made up of the various grades shown in Figure 2 (52). The remaining percentage of the wheat berry is classified as millfeed. The flour can be classified in several ways. Straight flour is all the flour produced, with various streams of flour mixed into one. Flour emerges at a number of points in the milling process, with the purified middlings yielding the extrashort or patent or bread flour. In hard wheat mills, as much as 75–80^o % may be run together as first clear or split into fancy clear and second clear.

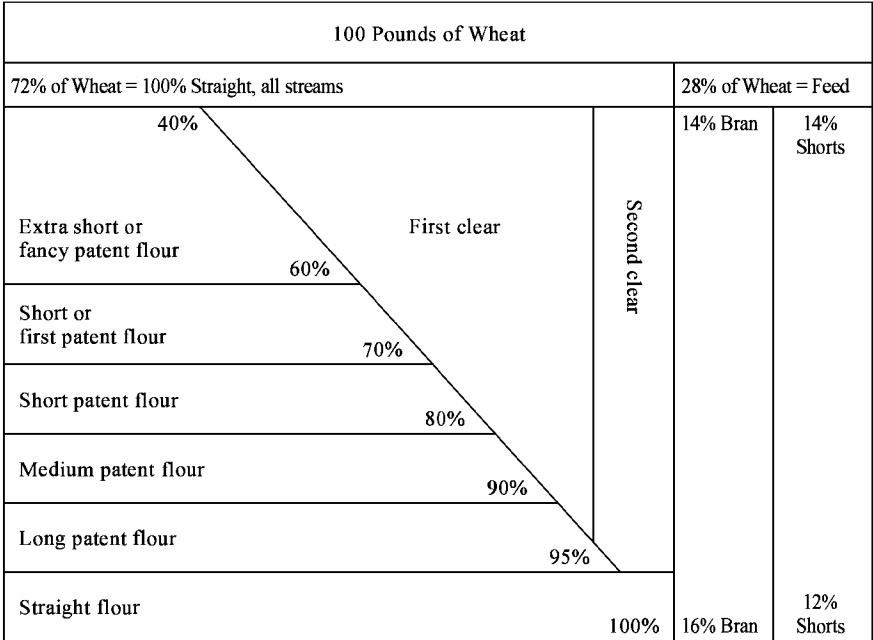


Fig. 2. Grades of flour. From Ref. 52.

In a soft wheat mill, 40–60° of the fancy patent may be taken off separately, leaving about 55° of the remaining flour to be classified as fancy clear. The chart (Fig. 2) is a generalization rather than an exact description of the yield of and single mill from a particular kind of wheat. It shows how the various streams of flour may be classified, starting with fancy patent, through short, medium, and long patent flours, leaving less and less to be classified as clear flour. The extrashort or fancy patent is the finest, with grades dropping down the scale to clears (53,54).

Toward the end of the millstream, the finished flour flows through a device that releases a bleaching–maturing agent in measured amounts. For bakery customers, the finished flour flows into hoppers for bulk storage, since most bakers add their own form of the enrichment formula to dough—a combination of thiamine, niacin, riboflavin, and iron. For packing as family flour, the enrichment ingredients are added in another mixing machine as the flour flows to the packing room. If the flour is self-rising, a leavening agent and salt are also added.

In milling of durum wheat for the macaroni trade, special equipment is required, especially additional purifiers to separate the bran from the semolina, a coarse granulation of the endosperm (55). By federal definition, semolina is prepared by grinding and bolting durum wheat, separating the bran and germ, to produce a granular product with no more than 3° flour. Durum millers also make granulars, or a coarse product with greater amounts of flour; or they grind the wheat into flour for special use in macaroni products, particularly in noodles.

Other Milling Developments. In the early 1960s it was found that wheat could be ground to a very fine flour by impact milling and that the product could be further separated into products varying widely in protein. Finished flour from a conventional roller mill is further reduced in particle size in a high speed grinder, an impact or pin mill. Disintegration of the flour particles takes place as they strike one another, the surface of the rotors, and the pins. The flour is fractured in granular form rather than pressed and broken as in a roller system. The reground flour contains a mixture of relatively coarse endosperm chunks, intermediate fragments, and fines.

The reground flour is channeled to a classifier, where swiding air funnels the larger particles down and away while the smaller fines are lifted up and separated. Repeating the process 20–30° of the flour is separated into a low protein product suitable for cakes and pastries; 5–15° makes up the fine fraction, containing 15–22° protein.

The high protein flour may be used to fortify or blend with other flours. Recombining some of the high protein fraction with the coarse portions permits a miller to tailor a flour of protein value to a buyer's specifications.

Fine grinding and air classification make possible the production of some cake flour from hard wheat and some bread flour or high-protein fractions from soft wheat. Application of the process theoretically frees the miller from dependence on different wheats, either hard or soft, that change each crop year. The problem is how to market the larger volume of low protein or starch fractions at prices adequate to justify the installation and operation of the special equipment (46).

Bleaching and Maturing Agents. Flour is often bleached, both for use by bakers and for the family flour trade. Bleaching improves the color of the flour, and some bleaching agents mature the flour and condition the gluten, improving baking quality. Flour fresh from the mill may be aged in storage or treated with a bleaching–maturing agent to improve baking quality. In aged, unbleached flour or in treated flour, the proteins are slightly oxidized by oxygen from air. The oxidation of flour makes gluten stronger or more elastic and produces better baking results. Cookie, pie crust, and cracker flours usually perform better when no bleach is used.

Dough handling properties can be modified beneficially, or adversely, by the addition of minute amounts of reducing agents (such as cysteine). In addition, the performance of a flour can be improved significantly by the addition of certain oxidizing agents. A list of oxidizing and reducing agents used to improve the baking quality of wheat flour is given in Table 3.

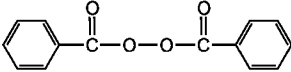
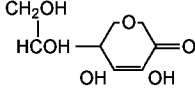
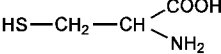
Table 3. Flour Treatment Agents^{a, b}

	Compound			
	Bromate	Iodate	Azodicarbonamide	Chlorine dioxide
structure	KBrO ₃	KIO ₃	$\text{H}_2\text{N}-\overset{\text{O}}{\parallel}{\text{C}}-\text{N}=\text{N}-\overset{\text{O}}{\parallel}{\text{C}}-\text{NH}_2$	ClO ₂
form	powder	powder	powder	gas
introduced	1916	1916	1962 patent	1948
regulations				
flour	50 ppm max U.S. whole wheat, 75 ppm max	not allowed	45 ppm max	sufficient for bleaching
bread	U.S.—75 ppm max (see below) Can.—100 ppm max (see	45 ppm max (see below)	U.S. and Canada, 45 ppm max	

labeling	below) potassium bromate or bromated flour	potassium iodate	bleached flour	bleached flour
used in action	bread flour and dough oxidant—improver	bread dough oxidant—improver, mix reducer	bread flour and dough oxidant—improver, maturing	bread flour oxidant—maturin g, bleaching
reaction rate	slow	fast	fast	fast
reaction product	bromides	iodides	biurea	chlorides
usage rates, ppm				
flour, normal	15	0	5	15
total, range	10–75	10–20	2–20	
overtreatment effects	open grain holeskeyholingincorrect volume	low volumeweak, sticky doughs	dry, bucky dough; low volume; poor grain; crust checking	poor volume; poor grain
testing				
flour, spot	AACC 48-02		no	no
flour quantitative	AACC 48-42		yes—difficult	no
bread	very difficult Bromolux	difficult	no Maturox	no Dyox
Pennwalt products	No—D-Lay		Dependox	

^a Courtesy: Pennwalt, Buffalo, NY.

^b These oxidating–reducing agents are used to improve the baking quality of wheat flour. They are added at the flour mill, the bakery, or both. Having the proper oxidation treatment is critical: it ranks second only to protein quality and quantity in determining flour quality. Government regulations on bakery use are as follows: United States—potassium bromate, calcium bromate, potassium iodate, calcium iodate, and calcium peroxide, or any combination of these, cannot exceed 75 ppm. This includes any bromate added to the flour. Canada—potassium bromate, calcium peroxide, ammonium persulfate and potassium persulfate, or any combination of these, cannot exceed 100 ppm. This includes any bromate added to the flour. Calcium iodate, potassium iodate or any combination of iodates cannot exceed 45 ppm.

Compound				
Chlorine	Benzoyl peroxide	L-Ascorbic acid	L-Cysteine	
Cl ₂				
gas 1912	powder 1921	powder 1935 200 ppm max	powder 1962 patent	
sufficient for bleaching	U.S.—sufficient for bleaching; Canada—150 ppm max	U.S.—no limit; Canada—200 ppm max	U.S.—not allowed; Canada—90 ppm max U.S. and Canada—90 ppm max	
bleached flour	bleached flour	ascorbic acid added as a dough conditioner	L-cysteine	
cake flour	bread and cake flour	bread flour and dough	flour (Canada only), miscellaneous doughs	
oxidant—maturing, bleaching	oxidant—bleaching	reductant—oxidant—mix reducer, improver	reductant—mix reducer, relaxer	
fast	1–3 days	medium	fast	
chlorides	benzoic acid			
120	50	70	30	
25–500	30–100	50–100	15–70	
low volume, dense crumb	none	none	weak, sticky doughs; poor grain and texture	
no	AACC 48-06B	yes—good	yes—fair	
pH used	flour color	yes	no	
no	benzoic acid—hard	no	no	
flour	Novadelox	No—D-Lay	No—D-Lay	
chlorination systems	Zerolux	AA-25	Assist	

Flour Types

Hard-wheat flours are usually higher in protein than are soft-wheat flours. They may be milled from either winter or spring wheat varieties. Those with highest protein content, characterized by their capacity to develop the strongest gluten, are used in commercial bread production where doughs must withstand the rigors of machine handling. Other hard-wheat flours with more mellow gluten, easier to develop in kneading by hand, are packed as family flour, all-purpose flour, and self-rising flour. The protein in hard spring wheat flour runs from 11 to 16° º; in hard winter wheat flour from 10–14 (47,48).

Soft-wheat flours are sold for general family use, as biscuit or cake flours, and for the commercial production of crackers, pretzels, cakes, cookies, and pastry. The protein in soft wheat flour runs from 7 to 10° º. There are differences in appearance, texture, and absorption capacity between hard- and soft-wheat flour subjected to the same milling procedures. Hard-wheat flour falls into separate particles if shaken in the hand whereas, soft-wheat flour tends to clump and hold its shape if pressed together. Hard-wheat flour feels slightly coarse and granular when rubbed between the fingers; soft-wheat flour feels soft and smooth. Hard-wheat flour absorbs more liquid than does soft-wheat flour. Consequently, many recipes recommend a variable measure of either flour or liquid to achieve a desired consistency.

Whole wheat flour, according to FDA specifications, is a coarse-textured flour ground from the entire wheat kernel. It contains the bran, germ, and

endosperm. The presence of bran reduces gluten development. Baked products made from whole wheat flour tend to be heavier and denser than those made from white flour. Whole wheat flour is rich in B-complex vitamins, vitamin E, fat, protein, and contains more trace minerals and dietary fiber than does white flour. In most recipes, whole wheat flour can be mixed half and half with white flour for a satisfactory product. Graham flour is synonymous with whole wheat flour, named after a physician, Dr. Sylvester Graham, who advocated the use of whole wheat flour in the first half of the nineteenth century (47,48).

According to FDA standards, wheat flour must be milled from cleaned wheat, essentially free of bran and germ, and be ground to such a degree that the product will pass through a sieve having 65 openings per lineal inch (or 4,225 openings per square inch). The moisture content must not exceed 15%.

All-purpose flours are designed for home baking of a wide range of products—yeast breads, quick breads, cakes, cookies, and pastries. Such flours may be made from low protein hard wheat, from soft or intermediate wheats, or from blends of hard and soft wheat designed to achieve mellow gluten a homemaker can manipulate, and tenderness of structure in the final baked product. All-purpose flour, sometimes called general-purpose or family flour, is most commonly used for home baking. Most modern recipes for home baking are designed for use with all-purpose flour.

Self-rising flours are all-purpose flours to which leavening agents and salt have been added. The leavening agents used are sodium bicarbonate and an acid-reacting substance, monocalcium phosphate, sodium acid pyrophosphate, sodium aluminum phosphate, or a combination of these acids. The sodium bicarbonate and the acid ingredient react in the presence of liquid to release carbon dioxide. The leavening agent and salt are added in amounts sufficient to produce not less than 0.5% of their combined weight of carbon dioxide. Their combined weight must not exceed 4.5 parts per hundred parts of flour. Phosphated flour is all-purpose flour to which the acid-reacting ingredient, monocalcium phosphate, has been added in a quantity of not more than 0.75% of the weight of the finished phosphated flour. It assists in stabilizing gluten and helps nourish yeast. Bromated flour contains potassium bromate in a quantity not exceeding 50 ppm. The bromate has an oxidizing effect that improves the baking qualities of the flour. Such flour must be labeled as bromated.

Cake flours are milled from low protein soft wheat especially suitable for baking cakes and pastries or from low-protein fractions derived in the milling process. Cake flours are usually not enriched, but are bleached. The sale of cake flours declined during the 1950s as packaged cake mixes became more popular. Instantized, instant blending, or quick-mixing flour is a granular or more dispersible type of product for home use. It is free-pouring like salt, and dust-free compared to regular flour. It eliminates the need for sifting, since it does not pack down in the package and since it pours right through a screen or sieve. Granular flour instantly disperses in cold liquid rather than balling or lumping as does regular flour. The granular flour is produced by special processes of grinding and bolting, or from regular flour subjected to a controlled amount of finely dispersed moisture that causes the flour to clump or agglomerate. It is then dried to a normal moisture level (47,48). Gluten flour is milled to have a high wheat gluten and a low starch content and is used primarily by bakers for dietetic breads, or mixing with other flours of a lower protein content.

Semolina is the coarsely ground endosperm of durum wheat. High in protein, it is used by U.S. and Italian manufacturers for high quality pasta products such as macaroni and spaghetti. In Africa and Latin America it is also used for a dish called couscous. Durum flour, a by-product in the production of semolina, is used to make commercial American noodles. Farina is the coarsely ground endosperm of hard wheats. It is the prime ingredient in many American breakfast cereals. It is also used by manufacturers for inexpensive pasta.

Additional basic wheat products are wheat berry (kernel), bulgar, cracked wheat, wheat germ, bran, and commercial cereals. See Ref. 56 for more information on wheat science technology.

Rice: Production and Consumption

Rice is grown in more than 100 countries and on every continent except Antarctica. In the world economy rice is an extremely important food, second only to wheat in total world production, and its yield per hectare exceeds that of wheat (57). Rice is the main staple food for more than half of the world's population and it accounts for one-third to one-half of the daily caloric intake in many Asian countries. It is also the major source of protein for the masses of Asian people. In many African and South American countries rice is rapidly becoming the staple food for much of the population.

World rice production has been increasing at approximately 7 million t yr since 1950. Land use for rice production increased from 103 million ha in 1950 to 146 million in 1978. Since 1978 the acreage has remained fairly constant. The increase in production since 1978 has been the result of higher yields per ha (increasing from 2.58 in 1978 to 3.16 million t ha in 1988).

The major rice-exporting countries are Thailand, United States, Pakistan, EC-12, China, Burma, and Australia.

The major net exporters in the countries of the European Community (EC-12) are Italy and Spain. The United States only produces about 2% of the world's rice crop. However, because most of the world's rice is consumed within the borders of the producing countries (Asia produces and consumes about 90% of the world's rice), the United States, which exports approximately half of its production, is one of the leading rice exporting countries.

While U.S. rice exports have remained fairly stable at about 20% of the world market in the 1970s and 1980s (range 14.7–29.2%), fundamental changes in the market have altered the destination and type of U.S. exports (58). The Middle East and subsaharan Africa replaced Asia as primary U.S. export markets in the early 1980s. Parboiled rice, demanded by growing markets in the Middle East, has increased its share of U.S. rice exports, while brown rice, once demanded in large quantities by South Korea, has declined. Long-grain rice increased its share of U.S. rice exports to about 67% in the 1980s, due to expansion of the Middle Eastern markets and declines in Asian markets. U.S. government-sponsored export programs play a major role in promoting rice exports.

Nearly all domestic production of U.S. rice is located in the midsouth and California. Arkansas is the leading rice producing state with 40.6% of the national total in 1988. Following Arkansas in order of production share are California (18.5%), Louisiana (15.1%), Texas (14.6%), Mississippi (8.6%), and Missouri (2.6%).

Almost all rice grown in the United States is produced from varieties developed by research centers located at Stuttgart, Ark.; Biggs, Calif.; Crowley, La.; and Beaumont, Tex. (59). Traditionally, rice varieties are classed as long-, medium-, or short-grain types. Responsible breeding programs in the United States have produced varieties for each grain type that are associated with specific cooking, eating, and processing qualities. The long-grain varieties cook dry and fluffy and the cooked grains tend to separate, whereas high quality medium- and short-grain types are more moist, chewy, and sticky. There is demand for all three grain types. Different cultural groups prefer different and varied characteristics in their home-cooked rice. Processors require different quality characteristics for use in various prepared and convenience food products.

The majority of the rice grown in the United States is the long-grain type (74.3%) and this is the predominant type grown in the mid-South. Medium-grain rice accounts for 23.6% of the total U.S. production and is grown mainly in California, although significant quantities are also grown in Louisiana and Arkansas. The short-grain varieties (2.1% of total production) are grown almost exclusively in California.

The three principal domestic uses for rice in 1988 were direct food (61%), processed food (18%), and beer (20%). The direct food use figure includes the conventional white milled rice plus specialty rice products (parboiled, precooked, aromatic, brown, and prepackaged mixes) shipped directly from the rice mills. The specialty products account for approximately one-fifth of the direct food use. Approximately two-thirds of the direct food use rice is ultimately distributed to consumers through retail outlets and one-third through food service outlets.

The use of rice is processed foods in the United States has almost doubled over the last decade. Breakfast cereals are the foremost consumer use and mainly medium-grain is used by cereal manufacturers. Other processed food uses include prepackaged mixes (about 80% long-grain), canned soup (long-grain), frozen dinners (long-grain), candy (broken and medium-grain), baby food (rice flour), rice cakes (primarily long-grain), and pet food (broken and rice flour). Rice flour is also used by domestic processors in such products as dusting powder for various fried foods and pizza, nonallergenic baking powder, and snack foods. Rice use for beer increased rapidly in the 1970s and has held fairly stable in the 1980s.

The per capita consumption of rice in the United States has doubled since 1960 to approximately 10 kg in 1989. Over the last decade U.S. rice consumption has benefited from a growing trend in U.S. diets away from high fat animal products and toward grain-based foods. Many health groups encourage use of the complex carbohydrates found in grain products such as rice. Also, increases in the Asian and Hispanic segments of the U.S.

population have contributed to greater rice consumption. Per capita consumption of rice by these ethnic groups often exceeds 50 kg. Development of new products and increased promotion of rice by the U.S. Rice Council have helped make rice more available and visible to a wider range of consumers. Many products emphasize rice's versatility as a healthy food that can be consumed as a breakfast cereal, side dish, dessert, or snack.

Processing of Rice

Rice processing involves harvesting, drying, storage, and milling. These operations vary considerably throughout the world. The most important consideration during rice processing is prevention of breakage of the endosperm because broken rice is worth about one-half the value of head (whole grain) rice.

Storage and Handling. Practices for storing rice without losses due to microorganisms, insects, rats, mice, and birds are well known in developed countries (60). However, good storage practices are not always followed. Estimates of storage losses range from 2.5% of production in the United States (61) to as high as 30% in some developing countries (62).

Milling. The purpose of rice milling is to remove the hulls and bran from harvested, dried rough rice. Shelling refers to the removal of the outer hull or shell from rough rice. The operation is conducted in machines that have been known by different names such as shellers, hullers, huskers, dehuskers, and decorticators. Similarly, the shells are also known as hull, husks, and chaff. The term *hulling* in most parts of the United States has the same meaning as shelling; however, in some areas hulling also refers to the removal of both hulls and bran. After removal of the hulls, the rice is called brown rice.

The brown rice is milled in a milling machine whereby all or most of the bran is removed to produce the white milled rice. This milling operation to remove the bran is sometimes called scouring or whitening. In polishing of the milled rice, traces of bran that may remain on the rice after milling are removed and the rice surface is given a smoother finish. Total milled rice includes both the head rice and the broken rice. Head rice or head yield refers to the milled whole (unbroken) rice grains. The broken rice is generally subdivided into three sizes: second heads, screenings, and brewer's rice.

It should also be understood that in the rice industry a rice mill may refer to a machine for removing bran or it may include a series of processing operations that, in general, consists of cleaning, shelling, milling, sizing, packaging, and several auxiliary operations. The objective of these rice mill operations is to produce a white, whole-grain product that is essentially free of bran and foreign matter and that contains a minimum of broken kernels.

In some rice growing areas rice milling is accomplished by very primitive methods such as pounding the rough rice in a wooden mortar and pestle followed by winnowing. At the other extreme are very modern methods where milling is accomplished in large, highly automated plants. Thus there is no typical rice mill. However, the modern processing of rice consists of essentially these steps: (1) cleaning the incoming rough rice, (2) shelling the rough rice, (3) milling to remove the bran from the brown rice, (4) grading the milled rice by length into whole grain and different sizes of broken, (5) mixing milled whole grain and broken to meet specifications of buyers, and (6) packaging.

Parboiling. A hydrothermic process, parboiling will greatly improve the milling quality of rice such that head yields will approach total yields (ie, zero breakage). Comprehensive quantitative studies on the effects of parboiling on the breakage of rice have been reported and it has been noted that kernel defects such as cracks, chalkiness, and incomplete grain filling are completely healed during the parboiling process (63,64). When properly dried, the rice kernels are resistant to mechanical breakage. The milling quality of parboiled rough rice is determined largely by drying conditions following parboiling rather than the previous history or condition of the rice. Consequently, for rough rice that is to be parboiled the optimum harvesting, drying, and storage conditions should be selected on some basis other than that of preserving the milling quality. For the same reasons, parboiling is an excellent means for salvaging rice whose milling quality has been inadvertently damaged by improper handling or processing. For more information on rice processing and uses, see Ref. 65.

Corn: Production and Economics

Every continent, except Antarctica, grows corn; 40% of the present world crop is produced in the United States. In the 1987–1988 crop year, 12 states (Iowa, Ill., Nebr., Minn., Ind., Ohio, Wis., Mo., S. Dak., Mich., Kans., and Tex. in order of production) produced 157.5 million metric tons (6.2 billion bushels) that was 88% of the United States and 36% of the world's crop (66). Yield is influenced by many factors, including climate, pest control, planting density, and fertilization. Yield in the United States has increased from about 1.5 metric tons hectare in the 1930s to about 7.5 metric tons hectare. In 1985, a test plot produced 23.2 metric tons hectare and yields approaching 40 metric tons hectare are considered possible; corn is the most productive of the principal food crops.

A crop of corn is always maturing somewhere in the world. It grows from north latitude 58° to south latitude 40° and from below sea level to altitudes of 4,000 meters. It is adapted to areas with fewer than 10 in. of rainfall and regions having more than 400 in. Early varieties, that have been adapted to cold climates, mature in as little as 60 days. Late varieties grown in the tropics need nearly a year to reach maturity. Corn can grow to as little as 60 cm in height or as tall as 6.5 meters. Ears of corn can be as small as your thumb for some popcorn varieties or as large as 60 cm as those grown in the Jala Valley of Mexico (67).

A century ago, all corn was harvested by hand and stored on the cob in drying cribs. Today, one machine that picks and shells in the field can do the work of 50 workers. After harvesting, kernels are dried to less than 15% moisture content to maintain grain quality and prevent long term storage spoilage. Drying must be done carefully to allow the use in wet milling (68,69).

As with most major agricultural commodities, corn economics are tied to governmental policy. These policies, designed to achieve domestic objectives, generally force local prices above free market levels. A worldwide network of grain exchanges provides instantaneous response to perceived changes in supply, demand, and harvest estimates throughout the growing season. Seasonally produced commodities like corn generally are least expensive at harvest and rise in price to compensate for storage charges.

Corn Use in Feed

Although corn was originally grown as food, the single largest use today is as feed for farm animals. In the United States, 80% of the domestic corn usage is for this purpose. Swine consume ca one-third; beef cattle eat one-fourth; and dairy cattle and poultry each account for about one-fifth of the feed corn. Corn is an important feed ingredient because it supplies the energy component and a large portion of the protein input to the animal's diet. It has been a dominant force in feed manufacture, consistent in quality and composition, and never presenting its customers with a shortage of supply.

Also, the co-products of the various industries that use corn to produce beverage alcohol, starch, corn sweeteners, corn oil, and dry milled products provide the concentrated protein and vitamin content of corn, making them valuable feed ingredients as well. Corn gluten feed is a source of additional protein in feeds compounded for beef cattle and dairy herds. The fiber content is readily converted by these ruminants. The xanthophylls in corn gluten meal provide the coloration in chicken and eggs so desired by many United States consumers. The distilling industry provides distillers dried grain while dry millers produce high fiber, high calorie hominy feed.

Wet Milling

The United States wet millers buy ca 15% of the corn used in the U.S. (worldwide, wet millers consume about 10% of the corn used). There are two dozen corn wet mills in the United States ranging in capacity from 600–10,000 metric tons day. Shelled corn is shipped to wet millers by truck, rail, or barge. After cleaning to remove coarse material, ie, cobs, and fines (broken corn, dust, etc), the corn is steeped in a sulfurous acid solution to soften the corn and render the starch granules separable from the protein matrix that envelopes them. About 7% of the kernel's dry substance is leached out during this step, forming protein-rich steep-water, a valuable feed ingredient and fermentation adjunct.

The softened kernels are coarsely ground (first grind) to release the germs. Because of their high oil content, the germs are lighter than the starch, protein, and fiber fractions and can be separated in hydrocyclones. The germs are washed free of remaining starch, dried, and the valuable corn oil is removed by expelling, solvent extraction, or a combination of both. The spent germs are a valuable feed ingredient.

The starch–protein–fiber slurry is subjected to an intense milling to release additional starch from the fiber. The fiber is then wet-screened from the starch–protein slurry, washed free of starch, and dried to form the major component of corn gluten feed—a valuable feed ingredient. The best fiber can be additionally purified to become corn bran, a dietary fiber ingredient that has been shown to lower serum cholesterol and triglycerides (70). The starch-protein slurry is separated into its component parts, again taking advantage of the density difference between the heavier starch and the lighter protein (gluten) particles. Separation is usually done with combinations of disk-nozzle centrifuges and banks of hydrocyclones.

The protein fraction is filtered and dried to become high (60°) protein content corn gluten meal. The starch slurry can be dewatered and dried to produce regular corn starch. Dry starch can be sold as is or heat treated in the presence of acid catalysts to produce dextrins. Or, it is chemically modified before dewatering and drying to produce modified starches used in food and industrial applications. Lastly, it can be hydrolyzed to produce corn sweeteners.

Starch. Corn starch is a principal ingredient in many food products, providing texture and consistency, as well as energy. More than half of the corn starch sold is used in industrial applications, primarily in paper, textile weaving, adhesives, and coatings. Starch is a polymer consisting of α -linked anhydroglucopyranose units. Two forms exist: amylose is an essentially linear molecule in which the anhydroglucopyranose units are linked almost exclusively via α -1,4 bonds. Amylopectin is a much larger, branched molecule (the mol wt is ca 1,000 times greater than amylose) (71). α -1,4 Linkages predominate but there is a significant number of α -1,6 linkages that result in the branched structure. Although the ratio of amylose to amylopectin is quite consistent in normal corn varieties, it varies considerably when starch is obtained from either waxy or high amylose varieties.

When heated in the presence of water to 62–72°C, normal starch granules swell, forming high viscosity pastes or gels. This process is called gelatinization. Starch from normal corns form characteristic firm, opaque gels because of the amylose fraction. The linear molecules align on cooling after gelatinization in a process called retrogradation, forming a thick, rubbery mass. The bushy amylopectin molecules in waxy starch cannot align to form such a mass, resulting in softer, translucent salve-like gels. High amylose starches are difficult to fully gelatinize and provide little viscosity unless cooked above the boiling point of water. These vastly different characteristics are further enhanced by modification of the native granules, resulting in starches with a wide range of properties for industrial and food applications (see STARCH).

Corn Starch-Based Sweeteners. Acid or enzyme catalysts can be used to break the linkages between the anhydroglucopyranose units in the starch molecule with the addition of a molecule of water at the break site. This process, called hydrolysis, produces a variety of corn-based sweeteners. The first sweetener from starch (arrowroot, not corn) was produced in Japan in the ninth century. By the nineteenth century, starch sugars were being produced in Europe and the United States. Because glucose is not as sweet as sucrose, products made from corn syrups are not dloyingly sweet, allowing delicate flavors to reach the palate. However, enzymatic isomerization of glucose to fructose produces high fructose corn syrups (HFCS) that are as sweet as sucrose syrups, thus allowing corn sweeteners to replace sugar in liquid applications (such as soft drinks).

To produce sweeteners from corn starch, the starch is gelatinized in the presence of a catalyst under conditions that promote hydrolysis. Acid is usually used to make slightly converted or low dextrose equivalent (DE) syrups. Enzymatic conversion, using thermostable alpha-amylases (for liquefaction), beta-amylases (for maltose production), and glucoamylase (for high glucose content) is widely practiced to produce a full range of saccharide compositions. After the desired degree of hydrolysis is achieved, insolubles are separated by centrifugation or filtration (or both). Soluble impurities are removed using activated carbon or ion exchange (either singly or in combination) before evaporation of the purified syrup to the desired solids concentration. Pure glucose is obtained by crystallization from highly converted starch syrups. Highly converted glucose syrups can also be enzymatically isomerized to 42° (dry basis) fructose content, significantly increasing their sweetness. Further increases in fructose content are possible using chromatographic enrichment (72). Pure fructose is produced by crystallization from syrups enriched to fructose contents above 90° (dry basis).

Corn Oil. The crude corn oil recovered from the germs consists of a mixture of triglycerides, free fatty acids, phospholipids, sterols, tocopherols, waxes, and pigments (73). Refining removes the substances that detract from the quality, resulting in a nearly pure (99°) triglyceride stream. The first refining step is degumming; 1–3° water is added and the dense, hydrated gums are removed by centrifugation. The degummed oil is then refined. Treatment with dilute (12–13 wt %) NaOH forms water soluble soaps of the free fatty acids, allowing centrifugal separation. An alternative physical refining process steam strips the volatile components, primarily free fatty acids.

Pigments are removed by sorption on bleaching clay that is then separated from the oil by filtration. The oil is then winterized by cooling to ca 4°C, precipitating the waxes that are then filtered from the oil. Winterization is not required if the oil is to be hydrogenated. Deodorization, a steam stripping process similar to physical refining, removes volatile impurities, resulting in an oil with lighter color and improved oxidative stability.

Corn oil's flavor, color, stability, retained clarity at refrigerator temperatures, polyunsaturated fatty acid composition, and vitamin E content make it a premium vegetable oil. The major uses are frying or salad applications (50°) and margarine formulations (35°).

Drying Milling of Corn

The United States dry millers buy ca 3° of the corn used in the U.S. There are 71 dry corn mills in 22 states in the United States (74). They tend to be smaller than wet mills; while the largest U.S. dry mill has a capacity of 1,750 metric tons day, most grind fewer than 250 metric tons day. About one-fifth of the corn ground is white corn, the remainder being yellow dent. Dry millers use three processes; tempering-degerming, stone-grinding or nondegerming, and alkali-cooking.

Food Uses for Corn

Corn is used directly and ingredients produced from corn are widely used for food in Asia, Africa, North and South America, and parts of the former USSR. Much of the corn consumed in the United States is in the form of ready-to-eat cereal.

Corn as corn flakes, sweet corn, corn as various types of flour and meal, popcorn, other snacks foods such as chips, and corn juice as sweeteners, corn used in fermentation for beer and in the production of alcohol, and corncobs and stalks used as carriers for various chemicals and medications, as fiber sources, and for the improvement of soil condition by plowing under stalks, are some of the uses for this versatile crop. See Ref. 75 for more information on corn.

BIBLIOGRAPHY

"Cereals" in *ECT* 1st ed., Vol. 3, pp. 591–634, by W. F. Geddes, University of Minnesota, and F. L. Dunlap, Wallace & Tiernan Co., Inc.; "Wheat" in *ECT* 2nd ed., Vol. 22, pp. 253–307, by Y. Pomeranz, U.S. Department of Agriculture, and M. M. MacMasters, Kansas State University; "Wheat and Other Cereal Grains", *ECT* 3rd ed., Vol. 24, pp. 522–549, by Olaf Mickelsen, Consultant.

1. W. R. Aykroyd and J. Doughty, *Wheat in Human Nutrition*, FAO Nutritional Studies, No. 23, Food and Agricultural Organization, Rome, 1970, p. 1.
2. J. Storck and W. D. Teague, *A History of Milling Flour for Man's Bread*, University of Minnesota Press, Minneapolis, Minn., 1952, p. 27.
3. W. J. Darby, P. Ghalioungiu, and L. Grivetti, *Food: The Gift of Osiris*, Vol. 2, Academic Press, New York, 1977, pp. 492, 493.
4. Ref. 3, p. 460.
5. H. A. Wallace and W. L. Brown, *Corn and Its Early Fathers*, Michigan State University Press, East Lansing, Mich., 1956, p. 35.
6. Ref. 5, p. 33.
7. Ref. 2, p. 37.
8. D. K. Mecham in Y. Pomeranz, ed., *Wheat, Chemistry and Technology*, American Association of Cereal Chemists, Inc., St. Paul, Minn., 1971, p. 395.
9. L. Zeleny in ref. 8, p. 34.
10. S. Davidson, R. Passmore, J. F. Brock, and A. S. Truswell, *Human Nutrition and Dietetics*, 7th ed., Churchill Livingstone, Edinburgh, 1979, p. 43.
11. R. R. Williams, *Toward the Conquest of Beriberi*, Harvard University Press, Cambridge, Mass., 1961, p. 38.

12. Ref. 11, p. 41.

13. K. Y. Guggenheim, *Nutrition and Nutritional Diseases*, D. C. Heath & Co., Lexington, Mass., 1981, p. 172.

14. C. A. Elvehjem, R. J. Madden, F. M. Strong, and D. W. Wooley, *J. Am. Chem. Soc.* **59**, 1767 (1937).

15. E. V. McCollum, *A History of Nutrition*, Houghton Mifflin Co., Boston, Mass., 1957, p. 306.

16. J. B. Mason, N. Gibson, and E. Kodicek, *Br. J. Nutr.* **30**, 297 (1973).

17. A. S. Prasad in A. S. Prasad, ed., *Trace Elements in Human Health and Disease*, Vol. 1, Academic Press, New York, 1976, pp. 1–20.

18. D. Oberleas and A. S. Prasad in ref. 17, pp. 155–162.

19. A. R. P. Walker, F. W. Fox, and J. T. Irving, *Biochem. J.* **42**, 452 (1948).

20. H. H. Mitchell, *Nutr. Rev.* **3**, 130 (1945).

21. A. O'Boyle, T. Watkins, and O. Mickelsen, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **41**, 394A (1982).

22. S. Bolourchi, J. S. Feurig, and O. Mickelsen, *Am. J. Clin. Nutr.* **21**, 836 (1968).

23. T. Addis, E. Barrett, L. J. Poo, and D. W. Yuen, *J. Clin. Invest.* **26**, 869 (1947).

24. D. D. Makdani, M. Ahmad, B. H. Selleck, D. R. Rovner, J. S. Feurig, and O. Mickelsen, to be published in *Am. J. Clin. Nutr.*

25. R. P. Heaney, *Clin. Obstet. Gynecol.* **19**, 791 (1976).

26. E. T. Jensen and E. Ilstergaard, *Am. J. Obstet. Gynecol.* **67**, 1094 (1954); D. C. Smith, *N. Eng. J. Med.* **293**, 1164 (1975); T. M. Mack and co-workers, *N. Eng. J. Med.* **294**, 1262 (1976).

27. E. C. Reifenstein and F. Albright, *J. Clin. Invest.* **26**, 24 (1947).

28. T. V. Sanchez, O. Mickelsen, A. G. Marsh, S. M. Garn, and G. H. Mayor, *Proceedings of the Fourth International Conference on Bone Measurement*, National Institute of Health Publication #80-1938, Public Health Services, U.S. Department of Health and Human Services, Washington, D.C., May 1980.

29. O. Mickelsen and D. D. Makdani, *Proceedings of the National Conference on Wheat Utilization Research*, ARS-NC-40, Agricultural Research Services, U.S. Department of Agriculture, Washington, D.C., 1975.

30. O. Mickelsen, D. D. Makdani, R. H. Cotton, S. T. Titcomb, J. C. Colmey, and R. Gatty, *Am. J. Clin. Nutr.* **32**, 1703 (1979).

31. *Pediatrics* **23**, 400 (1981).

32. J. H. Shaw, *J. Am. Med. Assoc.* **166**, 633 (1958).

33. B. E. Gustafsson and co-workers, *Acta Odontol. Scand.* **11**, 232 (1954).

34. B. H. Landis in *Products of the Corn Refining Industry in Food, Seminar Proceedings*, Corn Refiners Association, Washington, D.C., May 9, 1978, pp. 47–54.

35. J. S. Trier in M. H. Sleisenger and J. S. Fortran, eds., *Gastrointestinal Disease, Pathophysiology, Diagnosis, Management*, 2nd ed., Vol. II, W.B. Saunders Co., Philadelphia, Pa., 1978, pp. 1029–1053.

36. W. K. Dicke, H. A. Weijers, and J. H. van de Kamer, *Acta Paediat.* **42**, 34 (1953).

37. **U.S. Pat. 1,367,530** (1921), J. C. Baker (to Wallace and Tiernan Co.).

38. *Fed. Regist.* **13**, 6969 (Nov. 27, 1948).

39. H. E. Magee, *Mon. Bull. Minist. Health Public Health Lab. Ser.*, 205 (Sept. 1950).

40. J. C. Drummond and A. Wilbraham, *The Englishman's Food*, Jonathan Cape, London, 1939, pp. 298–299.

41. D. P. Burkitt in D. P. Burkitt and H. C. Trowell, eds., *Refined Carbohydrate Foods and Disease, Some Implications of Dietary Fibre*, Academic Press, London, 1975, pp. 3–20.

42. D. P. Burkitt and H. C. Trowel in ref. 41.

43. T. G. Kiehlm, J. W. Anderson, and K. Ward, *Am. J. Clin. Nutr.* **29**, 895 (1976).

44. D. J. A. Jenkins, A. R. Leeds, M. A. Gassull, B. Cochet, and K. G. M. M. Alberti, *Ann. Intern. Med.* **86**, 20 (1977).

45. M. Stasse-Wolthuis, *World Rev. Nutr. Diet.* **36**, 130 (1981).

46. *From Wheat to Flour*, Wheat Flour Institute, Chicago, Ill., 1965.

47. *From Flour to Bread*, Wheat Flour Institute, Chicago, Ill., 1971.

48. *Nature's Best Wheat Foods*, Wheat Foods Council, Washington, D.C., 1984.

49. W. G. Heid, *U.S. Wheat Industry*, Agricultural Economics Report No. 432, U.S. Department of Agricultural Economics, Statistics and Cooperative Service, Washington, D.C., 1979.

50. National Food Review, *The Wheat Grower* **8**(9), 28 (1985).

51. Y. Pomeranz, *American Scientist* **61**, 683–691 (1973).

52. C. O. Swanson, *Wheat and Flour Quality*, Burgess Publishing Co., Minneapolis, Minn., 1938.

53. W. T. Yamazaki and C. T. Greenwood, eds., *Soft Wheat: Production, Milling, Uses*, American Association of Cereal Chemists, St. Paul, Minn., 1981.

54. J. H. Scott, *Flour Milling Process*, 2nd ed., Chapman and Hall, London, 1951.

55. G. Fabriani and C. Lintas, eds., *Durum Wheat, Chemistry and Technology*, American Association of Cereal Chemists, St. Paul, Minn., 1988.

56. Y. Pomeranz, "Wheat Science Technology," in Y. H. Hui, ed., *Encyclopedia of Food Science and Technology*, Vol. 4, John Wiley & Sons, Inc., New York, 1991, pp. 2823–2835.

57. U.S. Dept. of Agriculture, *Agricultural Statistics*, USDA, Washington, D.C., 1988.

58. N. W. Childs, "The Changing Role of the United States in the World Rice Market," in U.S. Department of Agriculture, *Rice Situation and Outlook Yearbook*, Economic Research Service, **RS-55**, Washington, D.C., 1989.

59. B. D. Webb, "Rice Quality and Grades," in B. S. Luh, ed., *Rice Production and Utilization*, AVI Publishing Co., Inc., Westport, Conn., 1980.

60. J. F. Steffe, R. P. Singh, and G. E. Miller, "Harvest, Drying and Storage of Rough Rice," in Ref. 59.

61. J. E. Wimberly, "Storage Practices," in *Meeting of Experts on the Mechanization of Rice Production and Processing*, FAO, Rome, 1971, pp. 142–144.

62. R. M. Saunders, A. P. Mossman, T. Wasserman, and E. C. Beagle, *Rice Postharvest Losses in Developing Countries*, ARM-W-12, U.S. Department of Agriculture, Washington, D.C., 1980.

63. K. R. Bhattacharya and P. V. Subba Rao, "Processing Conditions and Milling Yield in Parboiling of Rice," *J. Agri. Food Chem.* **14**, 473–475 (1966).

64. K. R. Bhattacharya, *Cereal Chem.* **46**, 478–485 (1969).

65. J. I. Wadsworth, "Rice" in Y. H. Hui, ed., *Encyclopedia of Food Science and Technology*, Vol. 4, John Wiley & Sons, Inc., New York, 1991, pp. 2264–2279.

66. *1988 Corn Annual*, Corn Refiners Association, Inc., Washington, D.C., p. 21.

67. P. C. Mangelsdorf, *Corn, Its Origin Evolution and Improvement*, Belknap Press, Cambridge, Mass., 1974, p. 2.

68. R. B. Brown, G. N. Fulford, T. B. Daynard, A. G. Meiering, and L. Otten, *Cereal Chem.* **56** (6), 529–532 (1979).

69. R. B. Brown, G. N. Fulford, L. Otten, and T. B. Daynard, *Cereal Chem.* **58** (1), 75, 76 (1981).

70. L. Earll, J. M. Earll, S. Naujokaitis, S. Pyle, K. McFalls, and A. M. Altschul, *J. Amer. Dietetic Assoc.* **88** (8), 950–952 (1988).

71. J. J. M. Swinkels, *Starke* **37** (1), 3 (1985).

72. L. Hobbs, F. W. Schenck, J. E. Long, F. R. Katz, *Cereal Foods World* **31** (12), 852–869 (1986).

73. R. A. Reiners, "Corn Oil in Products of the Corn Refining Industry" in *Foods Seminar Proceedings*, Corn Refiners Association, Inc., Washington, D.C., 18–23 (1978).

74. J. M. Russo and V. Heiman, *Poultry Science* **38** (1) (1959).

75. F. W. Schenck, "Corn and Corn Products," in Y. H. Hui, ed., *Encyclopedia of Food Science and Technology*, Vol. 1, John Wiley & Sons, Inc., 1991, pp. 482–490.

WHEAT GERM OIL.

See FATS AND FATTY OILS.

WHEY.

See FEEDS AND FEED ADDITIVES.

WHISKEY.

See BEVERAGE SPIRITS, DISTILLED.

WHITE LEAD.

See PIGMENTS, INORGANIC.

WHITENING AGENTS.

See FLUORESCENT WHITENING AGENTS.

WHITEWARES.

See CERAMICS.

WHITING.

See PIGMENTS, INORGANIC.

WINE

The unmodified word *wine* signifies the juice of grapes, particularly wine grapes, fermented by yeast and appropriately (and legally) finished into an alcoholic beverage. Wines are made from other fruits in much smaller amounts and, particularly by home winemakers, from other food materials such as rhubarb. These products require modifiers for the term *wine*, eg, orange wine, blackberry wine, dandelion wine, whey wine, etc. Rice, being starchy, when fermented has more in common with beer, and its product should be considered sake instead of rice wine. Fruits other than wine grapes typically are too tart and too low in sugar to make satisfactory wine directly. They must be ameliorated (within legal limits) by dilution of the acid and addition of sugar.

Status

As commercial food products, wines and the industry that produces them are unique in many ways (1,2). Whereas a few very large producers account for a major fraction of U.S. production volume, there are many small producers who may grow their grapes, make them into wine, bottle it, and even sell much of it directly to retail customers. Their position is analogous to that of a shepherd who grows the pasture and tends the sheep, turns the wool into cloth, makes the suit, and markets it. This high degree of vertical integration also often applies in the largest of the winemaking operations, which may own glassworks, etc, as well as vineyards. Whereas wineries can be extensive operations employing specialized equipment, they may be successful, even today (ca 1997), if small, modestly equipped, and focused on one or a few types of wines. Owing to a short season of ripe fruit and the very limited possibility of adapting a winery to process other foods, much winery equipment has only a short and intensive period of annual use.

Standardization is usual in most products, but diversity is sought and admired in wines. There is no other product which has a similarly extensive aficionado's literature, including dedicated magazines, newspaper columns, newsletters, winery letters, etc, as well as specific scientific and technical journals on viticulture (grape growing) and enology (wine science and technology). Diversity is sought not only among classes of wine, eg, sparkling wines vs sherries, but more significantly among wines of the same class.

Some of these differences are related to the grape variety, to the geography of the vineyard, and to the weather of a particular year (vintage). Even if these are essentially similar, individual wines have cachets and nuances that are not always attributable to specific technical differences, but which nevertheless distinguish them to the discriminating consumer. Different winemakers and sometimes different lots by the same winemaker capitalize upon these subtle differences in style to maintain consumer interest. Each red table wine labeled as a single variety, of the same vintage, coming from similar area, for example, is almost certain to differ in style among producers. One winemaker may emphasize later picking of the grapes, another storage in new oak barrels, a third longer pomace (skins and seeds) contact with the fluid, etc. To retain satisfied customers, some consistency is sought, but differences among vintages and styles are important in sustaining interest in wine and the industry of winemaking.

As is discussed herein, wines retain an aspect of mystic speciality and connections to art and culture. Detailed connoisseurship and wine appreciation

are occasionally deplored as a form of snobbishness. They are, nevertheless, important in themselves and can prove entertaining to understand. Pure technology cannot fully displace the romance of wine. Today most of the world's wine is produced by modern technology and is improved thereby, but winemaking is still something of an art, as far as the attainment of its ultimate quality is concerned. Considerable experience and study are required to appreciate fully all the kinds of wine available; no one can be up-to-date on all labels.

Magnitude

Production of grapes is a very large endeavor in a number of countries and, notwithstanding the importance of table fruit and raisins, commercial grapes mostly go into wine. Total tonnage produced annually has been larger for grapes, worldwide, than any other commercial fruit. In fact, grape tonnage in recent history has been about double that of oranges or bananas, and more than all other common sweet fruits combined (disregarding tomatoes and other fruits considered vegetables).

Recent hectareage, grape tonnage, and kiloliters of wine produced worldwide are given in Table 1. Annual wine production of 26×10^9 L translates into about 35×10^9 bottles, or about 6 for every person. A decreasing vineyard area after 1980 may be noted. Owing to overproduction, major producing countries have instituted programs designed to decrease wine production by restricting wine vineyards. Grape tonnage has not fallen in proportion to the decrease in vineyard area, however, because the least valuable vineyards are removed first and the number of metric tons per hectare, t ha, has risen. Wine production has fallen in response to more of the grapes going into other uses and a higher proportion of white wines. These factors help explain why the apparent yield of wine per ton of grapes has fallen. About 0.58 kl t or more should be obtained for white table wine and as much as 0.73 kl t for red. It must be noted that such compilations depend on reports from different countries where reporting practices and reliability of data vary.

Table 1. World Grape Hectareage, Tonnage, and Wine Production^a

Period	Vineyard, ^b 10 ha	Grapes, ^b 10 t	Yield, ^b t ha	Wine total, ^b 10 kl	Wine yield, ^b kl t
1967–1970	9.9	52.1	5.3	28.4	0.55
1971–1975	10.0	55.4	5.5	31.3	0.56
1976–1980	10.2	60.6	5.9	32.6	0.54
1981–1985	9.8	62.4	6.4	33.3	0.53
1986–1990	8.8	60.6	6.9	30.2	0.50
1991	8.5	57.8	6.8	25.9	0.45
1992	8.4	60.9	7.2	29.3	0.48
1993	8.3	56.5	6.8	25.9	0.46

^a Ref. 4.

^b All amounts for 1967–1970, 1971–1975, 1976–1980, 1981–1985, and 1986–1990 shown are averages for the respective periods.

The percentage by country of world wine production is given in Table 2. Italy and France together produce about 45%. The United States and Australian percentages have been rising. Countries with the largest production, not surprisingly, have large consumption, usually both in total and per capita (Table 3). Decreasing consumption per capita has been quite general and prolonged. France and Italy consumed nearly 100 L person annually as recently as 1980. A glut of excess wines from such countries affects markets widely, because the excess drives prices of ordinary wine down and increases efforts to export.

Table 2. Percentage of the World's Wine by Selected Major Producing Countries^a

Country ^b	Year						
	1993	1992	1991	1990–1986	1985–1981	1980–1976	1975–1971
Italy	24.0	23.4	23.1	21.3	21.6	22.8	22.2
France	20.6	22.3	16.5	22.3	20.2	20.6	22.0
Spain	9.8	11.6	12.1	11.5	10.2	10.4	10.3
United States	6.6	5.3	5.9	5.8	5.3	5.1	4.2
Argentina	5.6	4.9	5.6	6.4	6.1	7.5	7.3
Germany	3.7	4.6	3.9	3.7	2.9	2.4	2.6
South Africa	3.5	3.4	3.8	2.9	2.6	1.9	1.7
Roumania	2.3	1.6	1.9	2.6	2.6	2.5	2.5
Australia	1.8	1.6	1.5	1.5	1.2	1.1	0.8
Portugal	1.7	2.6	3.8	2.9	2.7	2.9	3.3
Chile	1.3	1.1	1.1	1.4	2.0	1.7	1.6
all other	19.1	17.6	20.8	17.7	22.6	21.1	21.5

^a Ref. 4.

^b Data unavailable for the former USSR; declined since 1985–1981, when it made about 9%.

Table 3. Representative Consumption per Person^{a,b} L/yr

Country	Year						
	1993	1992	1991	1990	1985	1980	1970
France	65	66	67	73	80	95	109
Italy	63	61	60	71	82	93	100
Portugal	55	55	62	50	87	70	77
Argentina	48	52	55	54	60	76	92
Switzedand	44 ^c	44 ^c	47 ^c	47 ^c	48 ^c	47 ^c	39 ^c
Spain	39	39	40	47	48	65	62
Chile		30	30	30	40	47	44
Germany	23	23	26	26	25	26	16
Australia	18	19	18	18	21	17	9
New Zealand	11	13	12	13		11	5

Sweden	12	12	13	13		10	6
Great Britain	12	10	10	12	10	7	3
South Africa	8	9	9	9	10	9	11
United States	6.7	7.1	7.1	7.8	9.2	8.4	5.0
Japan	1.0	0.9	0.9	1.1		0.4	0.3

^a Refs. 4 and 6.

^b Ordinarily based upon total population, not just adults. A small fraction of the population usually accounts for a major portion of the consumption.

^c Tourists probably inflate the Swiss consumption figures.

The quantity of wine consumed as a fraction of that produced for several countries is given in Table 4. France, Italy, Spain, Australia, and South Africa divert and export a sizable fraction of their production. Countries with little or no production, exemplified here by Great Britain, and those like Germany which regularly consume more than they produce, are obviously of strong interest to the high producing, oversupplied countries. Because wine is often consumed more than a year after being produced, a bad year may result in atypically large proportionate consumption for some other vintages. Examples of this effect appear in Table 4, eg, Chile in 1991.

Table 4. Wine Consumed as a Fraction of That Produced^a

	Year				
	1993	1992	1991	1990–1986	1985–1981
France	0.71	0.58	0.89	0.63	0.68
Italy	0.55	0.51	0.59	0.62	0.64
Germany	1.87	1.83	1.54	1.43	1.62
Portugal	1.38	0.73	0.57	0.69	0.88
Roumania	0.97	0.96	1.42	0.88	0.79
Spain	0.60	0.45	0.74	0.51	0.58
U.S.A.	1.05	1.15	1.17	1.22	1.15
Argentina	1.09	1.17	1.23	0.95	0.99
Chile	0.69	0.75	1.82	0.84	0.78
S. Africa	0.37	0.35	0.26	0.39	0.36
Australia	0.69	0.71	0.77	0.74	0.75
Great Britain	376.00	231.00	243.00	556.00	831.00

^a Ref. 4.

The United States is unusual among wine-producing countries in having sizable production and low per capita consumption, yet high imports relative to exports. Consumption varies by region, ranging from California at nearly 11 L person⁻¹ yr (exceeded only by Washington D.C. and Nevada, augmented by visitors) to Utah and West Virginia at 2 L person⁻¹ yr. A high proportion of U.S. wine is made in California (Table 5), but most states have some commercial wineries. Many have special laws favoring locally grown wines.

Table 5. U.S. Wine Production and Importation^a

Year	Total, 10 L	California, ^o %	New York, ^o %	Washington State, ^o %	Other states, ^o %	Export, 10 L	Export, \$ L	Import, 10 L	Import, \$ L
1994	1448.1	89.6				120.9	1.48	265.8	3.88
1993	1435.5	92.1	3.8	1.5	2.6	129.9	1.38	246.0	3.99
1992	1537.8	91.4	3.4	2.0	3.2	140.4	1.24	269.0	4.06
1991	1535.9	92.3	3.4	1.1	3.2	117.3	1.25	231.3	3.96
1990	1672.7	91.0	5.8	0.7	2.5	99.6	1.26	254.4	3.63
1989	1697.3	89.8	6.1	1.7	2.4	82.9	1.18	287.5	3.25
1988	1780.7	90.8	5.6	1.1	2.5	64.1	1.33	306.7	3.11
1987	1834.3	88.8	7.3	1.1	2.8	44.9	1.35	364.5	2.79
1986	1810.6	90.7	5.9	1.0	2.4	27.5	1.27	412.9	2.50
1985	1679.0	90.8	6.0	0.6	2.6	24.2	1.15	520.2	1.95

^a Ref. 6.

Wines may be imported for reasons at either of two extremes: because they are relatively inexpensive, or because they are in high demand, and thus command a higher price. Those exported from the United States tend to be the former, and those imported, the latter, based on the data shown in Table 5. There is some market, even in countries that are overproducing wine, for wines with exotic labels. This tendency illustrates another way in which wines differ from other products. Information on and cuisine from foreign countries associated with their wines serve as the vehicles for a form of a cultural exchange, besides the strictly commercial one.

Although, owing to changes in ownership, etc, it is impossible to give a permanently accurate number, there are presently (ca 1997) about 940 active bonded wineries in California and about an equal number in other states. One California winery is generally conceded to be the largest in the world, shipping about 50 × 10⁶ cases (1 case = 12 bottles, 750 mL each) per year, or about 35^o % of the California total.

Grape and wine production, marketing, and serving worldwide are estimated to employ directly at least tens of millions of persons (there are 2 million grape growers in France alone), and grapes occupy nearly 1^o % of the total agricultural land, worldwide. Clearly, they are important economically, not only as both everyday and premium beverages, but also in their contribution to feelings of contentment and perceptions of material well-being and sophistication.

History

The important role wines and their production have had in ancient times as well as in more modern periods of history, particularly in relation to science and technology, deserves wider appreciation.

Winemaking extends back to unrecorded times, perhaps 8000 BC. It was certainly well developed by 3000 BC in ancient Egypt, Mesopotamia, and the other areas considered the cradle of Western civilization. At the beginning of recorded history, wines were described, their production portrayed, and their properties critically examined. By 2500 BC, the Egyptians had evolved hieroglyphics describing various types of wine. As part of the funerary goods in King Tutankhamon's 1339 BC burial (7), included were wine amphorae stamped with the region of growth, the estate, the vintage, and the winemaker's name.

One was 35 years old.

The progenitors of the present wine grapes still grow wild in the Near East. In the wild state, grapevines grow in trees and are propagated by seed-scattering birds. A sugar content higher in the ripe grape than in all the other common fruits adapted grapes especially to attract birds. It also adapted them to ferment into wine. Broken open, grapes and their associated yeast ferment spontaneously to wine. One can imagine a small amount of wine forming briefly in a depression in a rock under wild vines. A passing person would likely find it pleasant and be entranced by the euphoria it produced. The impetus to reproduce this phenomenon must have been very great, in fact and soon proved successful.

A case can be made that winemaking was a cause, not just the result, of the early transition from nomadic to specialized, sedentary agricultural civilization. Cereals can be grown in many places. The plants have a relatively short lifetime, but the ripe grain can be gleaned over a considerable period and is readily transported. Grapes, however, are perennials taking a number of years to develop good crops of fruit from new plantings. Vines need to be pruned and tended to fruit well and must be guarded against depredations by other people as well as by birds and animals, particularly during the short and intense ripening stage. The fruit and wine are perishable.

Compared to grain, wine is less easily portable. Although beer can be made from grain throughout the year, there is the involved but necessary process to hydrolyze the grain's starch before beer can even be made. The grape's sugar, in contrast, is directly fermentable to wine. Both wine and beer were very early beverages, but of the two wine deserves the greater consideration in assessing the origins of agriculture and associated civilization. It is the easiest to make, and was probably the first alcoholic beverage where grapes grew.

In addition to its ancient origins, part of the traditional mysticism about wine relates to its euphoric effect. Certainly this would have seemed magical in earliest times. It contributed to involvement of wine in religion, in rituals and in celebration. This fact today is still reflected in the special ritualistic place accorded wines.

Grape cultivation is very ancient. One knows this from the steps necessary to convert the wild to the domesticated vine as well as from early writings and drawings. The earliest Egyptian tomb paintings show all vines productive with sizable crops. Wild grapevines grow as separate male and female plants, only the latter bearing fruit. Spontaneous mutants are hermaphroditic males, but they usually bear only a small amount of fruit. There were many varieties of grapes described, especially by Roman times. A long period of human intervention would have been required to convert the dioecious wild European vine into a large number of distinguishable *Vitis vinifera* varieties all self-fertile and bearing good yields. The term "variety of grape" signifies a cultivar (cultivated variety) representing a vegetatively propagated selection with specific characteristics (including, perhaps, a distinguishable varietal flavor in its wines) such as Cabernet Sauvignon or Chardonnay (Table 8) analogous to Red Delicious or Granny Smith apples.

Seeds cannot be used to propagate a new vine, if the characteristics of the mother vine are to be maintained. Rather, a clonal vine must be grown by rooting a cutting from a cane of the mother vine. The grapevine is naturally adapted to such propagation. Canes touching the ground easily take root and can be transplanted. A new vine takes at least three years to produce an appreciable amount of fruit from small rootings. Furthermore, if managed by pruning, the vine produces more consistent crops of better quality fruit. Earliest depictions and writings, especially from Rome and Egypt, show a well-developed agriculture, utilizing specialized tools for pruning and otherwise managing vineyards. These and other considerations (8) make wine an important but neglected contributor to the development and maintenance of civilization.

Writings from Ancient Greece and Rome frequently mention wine as part of everyday life (9). The Old and New Testaments of the Bible contain many references to wine. The universality of knowledge about wine was used in them to make scriptural ideas clear to the common people. For example, Jesus' first miracle involved turning water into wine when the supply of wine ran out at a wedding feast. The account not only shows the importance of wine in the celebration, but also it tells that the guests were surprised the later wine was better than the first. Good wine was recognized but uncommon, and lesser or spoiled wine was served after the guests were less critical.

During the Middle (Dark) Ages, wine and winemaking along with other sacraments and knowledge were husbanded in cloisters and enclaves which maintained the pockets of sophistication that enabled the later flowering of the Renaissance (9,10). Wine became further associated with art, letters, religion, and culture, and it remains so.

The development of modern microbiology resulted from initial studies focusing on wines and their yeasts. Being relatively large compared to bacteria, yeasts were early observed microscopically. Their causative role in fermentation was shown by Pasteur, insofar as sterilization by heating killed them and prevented fermentation as long as they were not reintroduced. The investigation into the mildest conditions needed to achieve this prevention so as to minimize flavor changes, the procedure now called pasteurization, was originally developed for wine, not for milk. Pasteur's work on the so-called diseases of wine, ie, acetic or undesirable malolactic fermentation, etc, led directly to clarification by him and others of the microbiological nature of anthrax and other infectious diseases of humans and other animals.

Biochemistry resulted from the early elucidation of the pathway of enzymatic conversion of glucose to ethanol by yeasts and its relation to carbohydrate metabolism in animals. The word enzyme means "in yeast," and the earlier word ferment has an obvious connection. Partly because of the importance of wine and related products and partly because yeasts are relatively easily studied, yeasts and fermentation were important in early scientific development and still figure widely in studies of biochemical mechanisms, genetic control, cell characteristics, etc. Fermentation yeast was the first eukaryote to have its genome elucidated.

Wine studies were crucial in the development of organic chemistry as well as microbiology and biochemistry. Polarized light was known to be rotated by products from living systems, but not by apparently identical synthesized compounds. The molecular reasons for this were unknown. Tartaric acid [87-69-4] is the major acid of grapes, but is present in very few other plant sources. It occurs in the L (+) form. Racemic acid [133-37-9] was the optically inactive form produced by racemization of the natural acid in alkali, raceme being the term for the kind of stem a grape cluster has, derived from a Latin word for grape clusters. Pasteur found that when the racemic derivative was allowed to crystallize slowly, two mirror-image crystals were produced. Separated by hand, the two forms rotated polarized light in opposite directions. The crystal structure mirrored the asymmetry of the molecules of tartaric acid. From such studies the whole field of stereochemistry was derived. Pasteur became famous for these studies in 1847 at the age of 26. It is said that is was a good thing Pasteur worked under less than ideal conditions, because, if his laboratory temperature had exceeded about 20°C, separate crystals would not have formed.

More recently, studies of wine and beer have initiated techniques of statistically valid sensory analysis. Scientific studies involving wine continue in these areas, building on past discoveries. Natural phenols as desirable dietary components and monitors of storage and aging reactions are currently active fields. Viticultural research, as well as enological, continues to improve grapes and the wines made from them (11).

Classification of Types and Styles of Wine

In classifying wines, many parameters might be used: fruit (species, variety, or condition), composition (color, alcohol, sugar, acid, or carbon dioxide), fermentation procedure (carbonic maceration, alcoholic, aerobic yeast film, malolactic), regulations (taxes, etc), geography, climate, weather (vintage). The last three are considered important in determining prices and reputations for individual wines, but are not very helpful in wine group classification. All of the above, and legally permitted variations, as well as different producer and advertising descriptions contribute to an almost infinite number of kinds of wines.

An effort to categorize all the significant classes of wine in the general form of a dichotomous key follows. A number of all-or-none alternatives are useful here, eg, does or does not the wine have added flavors, added alcohol, red color, significant sweetness, appreciable carbon dioxide, oxidized character, or recognizable flavor from the grape variety. It must be borne in mind that there is considerable leeway among styles within each type of wine. Not every wine fits neatly into such a classification. Regulations, taxation levels, geographical specifics, and tradition impose some limits on wines available in or from a specific locale.

Most wines with <14% alcohol are classed as table wines because they are usually consumed with meals. Note that as used here, premium wines are included. In the countries of the European Union (EU), table wine means only ordinary or everyday wine. Sparkling wines are included in this group because producing the sparkle and retaining it during consumption of a bottle by few people necessitates a modest alcohol level. The "generous" group of

wines have had distilled grape spirits added in order to reach, usually, about 18° ethanol. Such dessert and aperitif wines are intended to be consumed after or before a meal. Sweet and otherwise strongly flavored wines have gravitated here because they are more stable and thus can be consumed over several sittings. Other pertinent notes follow the list.

- I. "Natural" wines, < 14% alcohol. Their nature and keeping qualities have traditionally depended heavily on complete fermentation and protection from air.
- A. Wines without obvious carbon dioxide
- (1) Wines containing anthocyanin (red) pigments

a) Pink wines (can be made from most red grapes, by limiting extraction).

(1) Dry: ros  types including Tavel and varietals such as Gamay and Grenache

(2) Sweet: sweet ros s, "white" Zinfandel, etc

b) Full red color

(1) Dry (<0.5% sugar).

(a) Having usually distinguishable varietal aromas: Barbera, Barolo, Beaujolais, Bordeaux (M doc, St.  milion, etc), Burgundy, Cabernet Sauvignon, Ch teau-neuf-du-Pape, Chelo s, Chianti, Concord, Gamay, Hermitage, Petite Syrah, Pinot noir, Pinotage, Rioja, Ruby Cabernet, Syrah (Shiraz), Zinfandel, etc

(b) Without normally distinguishable varietal aromas: California (etc) dry red table wine (burgundy, claret, chianti), Carignane, Charbono, Cinsaut, Malvoisie, Mourastel, Valdepenas.

(2) Sweet. California (etc) red table and sweet red table, proprietary types; generally without specific fruit or varietal aroma unless Concord-type or fruit wines (blackberry, raspberry, strawberry, etc)

(2) White wines not containing anthocyanin (red) pigments

a) With (usually) distinguishable varietal aromas

(1) Not containing obvious sugar: Catawba, Chablis, Chardonnay, Chenin blanc, Delaware, Gew rztraminer, Graves, Moselle, Muscadet, Pinot blanc, Pinot gris, Rhine (Rheingau, etc.), Sauvignon blanc, S millon, Seyval blanc, Viognier, White Riesling, etc

(2) Containing considerable sugar: "Late Harvest" types of California (etc), German and similar (Auslese, Beerenauslese, Trockenbeerenauslese), Sauternes, Hungarian Tokay, sweet Catawba, light sweet muscat, sweet Sauvignon blanc, sweet S millon, and various proprietary wines

b) Without (usually) distinguishable varietal aromas

(1) Not containing obvious sugar: California (etc) dry white table wines (chablis, rhine, etc), Buarger, Colombard, Ugni blanc (Trebbiano), etc

(2) Containing considerable sugar: California (etc) sweet white table wines ("chateau" types, etc), various proprietarily labeled types

(3) Specialty products (partial list)

a) Flavored table wines: may be white, amber, tawny, or red. Flavors must be from a natural source such as a spice or fruit juices (no single other fruit can be imitated). Formula must be approved by U.S. Bureau of Alcohol, Tobacco and Firearms (BATF): proprietary brands. See also category III. B.

b) Nouveau-styles: red, prepared by carbonic maceration of the grapes, giving a special flavor; rushed to market: Beaujolais nouveau, California (etc) nouveau

c) Coolers, refreshment wines: made from white or red wine according to a proprietary formula. Low alcohol, often because of blending with various fruit juices: California Cooler brand and others.

B. Wines with obvious carbon dioxide

(1) From fermentation of added sugar; usual gauge pressure about 2.026 6.078   10  Pa (2 6 atm)

a) Containing anthocyanins and related (red) pigments

(1) Pink: sparkling ros , pink champagnes

(2) Red: sparkling burgundy, champagne rouge, Cold Duck

b) Not containing red pigments

(1) With muscat flavor: moscato spumante, sparkling muscat

(2) Without muscat flavor: Champagne, California (etc) champagne (bottle or bulk refermented), Cava (Spain), Sekt (Germany), espumante, shampanski, spumante, sparkling Catawba, etc

(a) Below about 1.5  sugar: brut style

(b) Above 1.5  sugar: (increasing) sec, demi-sec, doux

(2) Wines with excess carbon dioxide, not from refermentation of added sugar; Usual gauge pressure about 0.2026 2.026   10  Pa (0.2 2.0 atm)

a) Gassiness from fermentation of residual grape sugar: occasional wines from several countries sometimes called pearl or perle wines, California moscato amabile type

b) Gassiness from malolactic fermentation:inhos verdes wines (from Portugal, white and red)

(3) Carbonated wines: relatively rare; may be white or red

II. Wines with 14 17  ethanol

A. Miscellaneous, white and red, usually sweet types, mainly with proprietary names

B. Special types

(1) Blending: to increase the alcohol (or other component) or to impart a special character to other wines or even other products, eg, whiskey

(2) Sacramental, kosher, ecclesiastical, etc: usually sweet, eg, *vino santo*, with dessert-type names produces for special markets

(3) Refermented with aerobic yeast: light (white) flor (Fino) sherries, Spanish manzanilla types, California (etc) submerged culture dry sherry. See

Kirk-Othmer Encyclopedia of Chemical Technology (4th Edition)

367

also III.A.2.b) (1) (b) i)

III. "Generous" or Fortified Wines, 17–21° o alcohol. Their nature and keeping qualities depend heavily on the addition of distilled wine spirits.

A. Without added food flavoring materials

(1) Containing anthocyanins and related (red) pigments

- a) With a muscat flavor: certain pink and "black" muscatels from Aleatico or Muscat Hamburg, Mavrodaphne, etc
- b) Without a muscat flavor: ruby, tawny, vintage, port-type wines

(2) Not containing red pigments

- a) With a muscat flavor: Australian and California (etc) muscatel, Malaga (also raisiny), Muscat blanc (Muscat Frontignan, Muscat Canelli), Samos, Setubal, Sitges, etc
- b) Without a muscat flavor

(1) With a special (usually oxidized or heated) odor, the result of treatment or aging

(a) Without film yeast involvement

- i) Raisin-like flavor: usually an overlaid flavor, see 2a)
- ii) Baked odor: Madeira (usually sweet), California (etc) baked sherry (dry, medium, or sweet)
- iii) Cooked down must odor (caramelized): Marsala
- iv) Aged, oxidized, or rancio odor: Oloroso sherries, Banyuls (possibly tawny color from some red grapes), Tarragona (Priorato), etc

(b) With film (aerobic) yeast involvement

- i) Dry (<1% sugar): Spanish (flor or fino, amontillado, etc) sherries, various cocktail sherry-types from Australia, California (etc, including aerobic submerged cultured), Cyprus, South Africa, and the former USSR, Chateau Chalon (France). See also II.B.3.
- ii) Sweet: California medium or sweet flor sherry, some sweet Spanish types

(2) Without a special odor resulting from treatment or aging

- (a) With amber color: Angelica (California), white port (Portugal), etc
- (b) Without amber color: California white port

B. With added (natural) food flavors or plant extracts

(1) With a red color

- a) Proprietary types: includes wines containing quinine and similar additives (Byrrrh, Campari, Dubonnet, etc)
- b) Medicinal or homemade types: iron-containing wines, etc

(2) Without a red color

- a) Nearly dry: dry or French-type vermouth
- b) Sweet (usually with a muscat flavor): sweet or Italian-type vermouth
- c) Proprietary types: Mixed fruit-flavored (Thunderbird, etc). See also category I.A.3.a)
- d) Medicinal or homemade types: dandelion, gentian, rhubarb, etc

The "etc" following California usually represents other U.S. wines, such as New York or Oregon. Capitalized, Burgundy (and Champagne, etc) refers to a specific region of (in these cases) France. California burgundy (or champagne) uncapitalized refers to a wine considered a generic type produced in California (etc). By law the "California" must be on the label in substantially as large type as burgundy (or champagne, etc), so there is no chance for customers to be misled.

Two other general observations seem in order. First, proprietary wines are those whose characteristics are held by a specific producer as an owned secret and which are marketed as a brand-named wines. Their formulas must be approved by the BATF, if flavors are added. Second, many wines are blends, and particularly those bottled by a shipper can be much better than some of the individual wines in the blend. The more evocative French term, *négociant-à-leveur*, can be translated as merchant-rearer or raiser. Traditionally these shippers gathered wines from many small producers, usually farm winemakers, and improved them not only by blending, but also by further processing and maturation. Both of these two categories of wine, proprietary and shipper's, are difficult to include in a classification, and the list largely does not specifically account for these and styles within classes (types).

Composition

Clearly, with so many types (eg, dry red table wine) and styles (eg, oaky or not) within a type of wine, composition is quite variable, but table wines have certain consistencies in constituents and sweet wines overall tend to be similar, with some allowance given for their extra sugar and perhaps alcohol in certain cases. Components tend to be in two categories: those produced by fermentation (ethanol is a prime example) and those from the grapes (eg, red anthocyanin pigments). Of course, the constituents characteristic of a grape vary by variety and climatic and soil conditions, and they are modified by winemaking. It is seldom that the wine's varietal aroma is fully recognizable in the grape, for example.

A rough comparison of typical compositions of wines without unusual additions is given in Table 6. The apparent simplicity is, however, deceptive. Extensive listings of several hundreds of identified wine constituents are available (13–17), of which the data in Table 6 are only a useful summary. Each group of chemical constituents usually includes several, sometimes many, individual substances. As groups, sugars contribute sweetness and "body" and acids tartness. These are important sensory properties, but are relatively obvious and qualitatively if not quantitatively similar among compounds of the group. Together they account for much of the extract (nonvolatile solids) of wines. Glycerol is produced by fermentation, along with ethanol and a number of the volatile substances, especially higher alcohols and certain esters. These contribute to the typical vinous flavors of wine, but are rather similar among wines. A number of the other constituents, eg, pectins, ash, vitamins, fats, amino acids, and proteins, may be useful in comparing wines analytically, but do not ordinarily contribute to recognized sensory distinctions among wines. For example, detailed analysis of the metallic cations has been used to group wines by locality because the trace elements tend to reflect the soil without contributing directly to flavor.

Table 6. Estimates of Typical Gross Composition of Wines,^a wt %

	Table wines	Dessert wines
--	-------------	---------------

Component	White	Red	White	Red
water	87	87	76	74
ethanol	10	10	14	14
other volatiles	0.04	0.04	0.05	0.05
extract	2.6	2.7	10.1	12.2
sugar	0.05	0.05	8	10
pectin and related substances	0.3	0.3	0.25	0.25
glycerol and related substances	1.1	1.1	0.9	0.9
acids	0.7	0.6	0.5	0.5
ash	0.2	0.2	0.2	0.2
phenols	0.01	0.2	0.01	0.1
amino acids and related substances	0.25	0.25	0.2	0.2
fats, terpenoids	0.01	0.02	0.01	0.02
miscellaneous and vitamins	0.01	0.01	0.01	0.01
<i>Total</i>	<i>100</i>	<i>100</i>	<i>100</i>	<i>100</i>

^a Ref. 12.

Most of the sensory differences among wines, other than sweetness and acidity, are attributable to certain volatiles, terpenoids, and phenols. Rarely is a single volatile compound responsible for differences characteristic of a grape variety or wine type. Linalool [78-70-6] and other terpenoids and norisoprenoids account for the distinctiveness of muscat-aroma varieties and contribute heavily to Riesling, Gewürztraminer, and related varieties. Pyrazine derivatives, notably 2-methoxy-3-isobutylpyrazine [24683-00-9], produce the herbaceous or bell-pepper-like flavor found in varieties of the Cabernet and Sauvignon blanc groups. Methyl anthranilate [134-20-3] is strong in concord and related varieties, although in this and other examples these significant top-notes do not fully account for their complex aromas. As a rule, especially in the case of the more subtle flavors, wine aroma (grape odors) and bouquet (processing and aging odors) depend on a characteristic mixture of sensorily active components and cannot yet be fully explained.

Phenols make up an interesting group (14–16). Red wines are notably higher in phenol content (Table 6). This is partly a result of the fact that the red pigments themselves, anthocyanins, are flavonoids. In making such wines, the skins (and seeds) of the grape are more thoroughly extracted so that not only the anthocyanins of red grapes are extracted, but also other phenols including astringent tannins, bitter catechins, and oxidation-browning substrates. White wines from immediately clarified juice will have 100 mg L of phenols or so, mostly caftaric (caffeoyl tartaric) acid. Pale pink wines will add to that a few mg L of anthocyanin and young, tannic red wines perhaps a g L each of anthocyanins and other flavonoids including condensed tannins. Wines differ both qualitatively and quantitatively in phenolic components.

The phenols are genetically regulated in the grape and are different in presence or absence and relative amounts in different varieties, the most obvious example being white vs red grapes. Among red varieties there are many different patterns of the glycosides of malvidin [643-84-5], petunidin [1429-30-7], delphinidin [528-53-0], peonidin [134-01-0], and cyanidin [528-58-5]. Pelargonidin [134-04-03], important in strawberries, is absent in grapes. The anthocyanins are the 3-glucosides of these anthocyanidins in the European varieties, and in those derived from other *Vitis*, also the 3,5-diglucosides. The ratios vary, and although most varieties have portions of their anthocyanins acylated with acetic or hydroxycinnamic acids, some lack acylation. Unless they are degraded to smaller fractions, these compounds are not volatile, but are crucial to red color and, together with other natural phenols, to astringency, bitterness, and oxidizability of wines. Also, golden and amber colors of most white wines result from phenol oxidation and reaction.

Volatiles account for the aromas and bouquets of wines. These include some already mentioned plus other groups such as esters. Esters account for part of the fruitiness of a fresh young wine and its varietal distinctiveness, if it has any. Some of these esters and other odorants are present in the grape, others are made or released from grape components during winemaking, and still others (or portions of the same ones) are produced by the fermentation. Owing to the low pH and high ethanol content but limited amounts of the acids making up many esters in wines, there is major loss of many esters and great increase in others. *n*-Hexyl acetate [142-92-7] is important in fresh fermentation bouquet of wines from white juice fermented at cool temperatures. It is lost in relatively few days at room temperature because the equilibrium is too far toward hydrolysis and its pear-like note disappears. In contrast, monoethyl tartrate [608-89-9] forms slowly as part of aging owing to high content of both tartaric acid and ethanol. Whereas this ester is not volatile or odorous, a sufficient amount of it forms to lower the free acid and the tartness of the aging wine, thus mellowing it.

Chemical and Physical Analyses

Many analyses have been applied to wines, ranging from simple Brix and acid to detailed gas–liquid and high performance liquid chromatography, capillary electrophoresis, etc (16,18). Some are routine and needed by all wineries, and some are reserved for special situations. Highly sensitive methods and perhaps selective concentration are often required because low amounts may have importance and mixtures are complex. New techniques are continually being developed and applied. Because many techniques have special standing with regulatory agencies, a new procedure may be slow to replace another. Those interested in regulatory aspects should consult the publications of the Association of Official Analytical Chemists International in the United States, or the Office International de la vigne et du Vin in Europe.

Sensory Analysis, Wine Appreciation, and Spoilage

There is no combination of chemical or physical analyses which can, or is ever likely to, replace human sensory evaluation completely. Sensory examination of wines employs two major approaches: detecting differences and evaluating quality or, more briefly, analytical and hedonic (16,19). The former can be objective and the latter is inevitably somewhat subjective regardless of the expertise of the judges.

Almost everyone has sufficient sensory ability to become an expert wine taster. Most can develop the knowledge about wine, concentration, and sensory memory needed by an expert. Regardless of expertise, however, one person is not always able to reach the same conclusion, even in repeat tastings of the same wines. Experts often disagree in ratings in spite of diligent efforts to judge according to a standard for the wines in question. For these reasons, a panel of judges is needed and, if possible, replicate trials by every judge.

Various methods have been developed to eliminate biases which otherwise can skew results. The wines must be presented without identification, although the taster should be told the type of wines (the best strawberry wine should rate very poorly in a Cabernet class). For the most informative results, many details of coding, presentation order, replication, etc must be considered. The results must be statistically examined to estimate whether or not they could have been obtained accidentally. Statistical analysis is an entire field in and of itself, and wine studies have contributed greatly to its present sophistication, as applied in the flavor field.

The simplest analytical tasting can be a paired test: is the wine coded 67 more or less tart than that coded 19? Of course, the acids present and the pH can define chemical acidity. Sensory examination is still necessary to show whether or not an average person can detect the difference and how other tastes such as sweetness influence that detection. The paired test has the disadvantage that a leading question must be asked ("more acid?"). The next simplest analytical tasting method is the triangular test, which does not require such a question. Three samples are presented, two identical. The judge is asked to select the one of the three which differs from the other two. There is a right answer.

By such methods, the threshold (minimum) concentration to produce a detectable sensory difference in wines for an individual can be determined. By modifications of the questions asked, a recognition threshold concentration can also be determined. For example, a process producing a faint trace of extra acetic acid might make a wine detectably different from the untreated wine, but not certainly worse. At the point acetic acid is recognized as such, the wine would be downgraded as vinegary. To make such thresholds meaningful in relation to the general population, again, a sizable panel of individuals, replicated tests, and statistical validation are necessary.

For hedonic tasting, the judges and the panel manager should be clear on the question. Is it "How much do you like it?" or "How do you rank these wines?" or "How do these wines compare with a perceived ideal wine?" For the first two questions anyone can answer for themselves, but the answer must be replicated by several individuals, if it is to have general meaning. If the judges on the panel are able experts, the third question can be addressed. Even in such situations, as at well-run wine-fair judgments, at least three judges are used. Discussion is not permitted in analytical or research tastings. Neither is it at fairs until after separate examination, but it may be allowed before final awards in case one judge failed to note something the others did. Regardless of expertise and experience, for reasons already outlined, a single judge's evaluation is suspect and opinions can be influenced by the circumstances.

Descriptive analysis is a tool to standardize wine aroma terminology. The underlying idea of such analysis is to characterize a wine in terms familiar from other common sources such as fruits (cherry, raspberry, etc), vegetables (bell peppers, asparagus, etc), flowers (violet, rose, etc), flavors (cloves, mint, vanilla, etc), and others (hay, fresh-cut grass, etc). Defects can be noted by comparison with characteristic chemicals like ethyl acetate (part of vinegar character) or hydrogen sulfide (a reduction product). A wide selection of descriptors appropriate for wines has been organized into a "Wine Aroma Wheel" (16) that can help a winemaker, judge, or consumer to recognize, remember, and discuss an individual wine as well as to distinguish among different varieties, regions, and styles. Combined with carefully designed panel studies and multivariate statistical analyses such as principal component analysis (PCA) of the results, descriptive analysis is becoming an important tool, for example, to set norms for a particular type of wine and detect those that fall outside that norm.

Few commercial wines today are frankly spoiled. It must be noted that spoilage of wine does not pose a health risk to consumers. Because some of the spoilage odors and flavors are not easily analyzed chemically, sensory evaluation is the final judge. What is spoilage in one wine may not be in another. Yeast growth as a surface film on wine in the cask soon produces serious defects in other wines, but is required for the production of the aldehydes of flor sherries. If it is present at a recognizable level, unless it is part of the particular wine's nature and contributes to its attractiveness, a flavor is considered "off" unless adopted by the producers as part of their "style." There is a spoilage yeast that produces an odor in wines reminiscent of *para*-cresol [106-44-5] resulting from ethyl analogues. Considered as spoilage it is a defect, and many people do not like the odor, but a few localities and winemakers make a virtue of it. Disagreements among judges are a result.

Healthfulness

Recent recognition of healthy effects of moderate wine consumption has been overdue. Pasteur is often quoted to the effect that wine is the most healthful and hygienic of beverages. This is correct from the viewpoint that wine cannot be a vehicle for food poisoning. Even addition of typhoid bacteria to wine rapidly results in their death. None of the vectors of human disease and food toxicity can propagate or even survive in wine because of its low pH, alcohol content, often high tannin, etc. This is less crucial in modern societies, but is thought to be the justification of the earlier practice of giving babies water mixed with wine. The water was often contaminated.

This may also have been a factor in the conquering of the known world by the Roman Legions. Expeditionary and seige forces, even today but much more so then, are at risk for water-borne diseases. The practice of the Roman army to carry with it substantial wine supplies is seen as sound militarily for health reasons, rather than just from the standpoint of the enjoyment of the soldiers. Recent study indicates that tourist-type diarrheas are less frequently encountered if wine is consumed rather than water, even bottled water (20).

Owing to the residua of the Puritan Ethic and Prohibition as well as the undoubted contribution of overconsumption to the evils of drunk driving and alcoholism, alcoholic beverages have been seen as uniformly bad by health specialists. All the effort has gone into proving how bad they are and eliminating them. Wine was considered somewhat less dangerous because it was fairly dilute and consumed with meals, but most did not credit wine or any other such beverage with desirable effects. Ethanol is readily metabolized by the body (21), and any alcoholic beverage can have favorable effects, when properly consumed.

Only very recently have the government and other health agencies bowed to the preponderance of clinical, experimental, epidemiological, and historical evidence that moderate consumption of wine is not only not detrimental, but is beneficial. The proven benefit is in lowered incidence of cardiovascular complications in wine consumers. This apparently accounts for the relative health in this regard of people in France, Italy, and other Mediterranean countries even though their diet is otherwise less healthful (more fat particularly).

The 1995 *Dietary Guidelines for Americans* (22) acknowledges that moderate drinking of alcoholic beverages with meals is associated with a lower risk for coronary heart disease, the leading cause of death in this country. Moderation is defined very conservatively as one drink per day with a meal for a woman and two for a man (150 mL of table wine per drink). A flurry of recent studies further suggests that wine, particularly red table wine, has an additional favorable effect over other alcoholic beverages. This is attributed to the antioxidant, free-radical chain-breaking effect of the wine's natural phenolic compounds. Moderate drinkers' average longevity exceeds that of matched teetotalers, and both greatly exceed heavy drinkers in long and healthy life. In addition to direct health benefits, moderate consumption of wines with meals vs consumption at either extreme or consumption of other drinks is believed to promote conviviality and social health.

Technology

Ramifications of production of wine and related products are depicted in Figure 1 (23). Certain operations are required, unique, and irreversible in order to produce certain types of wine, eg, oxidation for sherries. Other operations such as clarification and tartrate stabilization are similar for all wines. Considering the complexity of the classes of wine already described, only general descriptions can be given (23,24). Because they represent such a large portion of total U.S. wine production (Tables 5 and 7), table wines and the practices in California are emphasized.

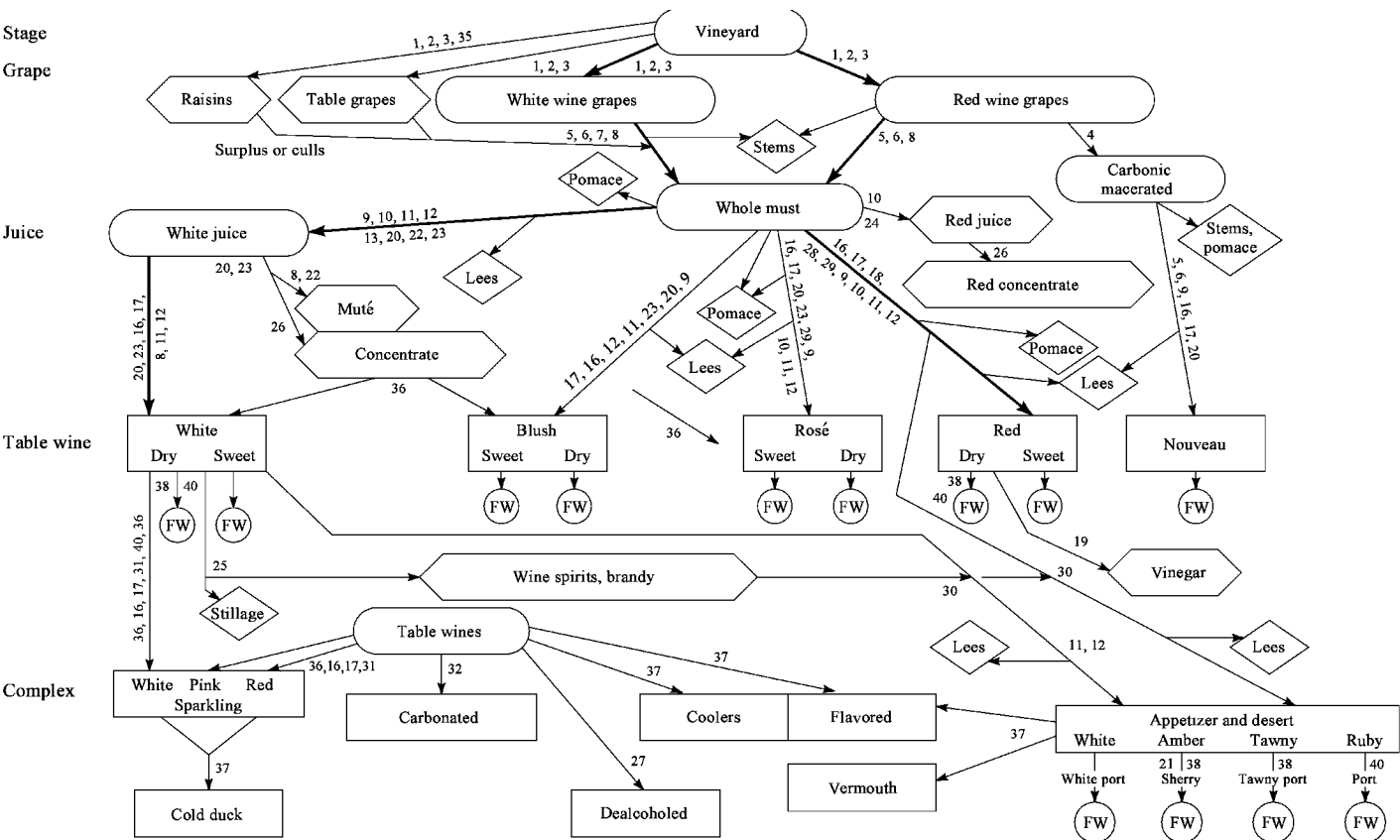


Fig. 1. An amplified outline scheme of the making of various wines, alternative products, by-products, and associated wastes (23). Ovals raw materials, sources; rectangles wines; hexagon alternative products (decreasing wine yield); diamond wastes. To avoid some complexities, eg, all the wine vinegar and all carbonic maceration are indicated as red. This is usual, but not necessarily true. Similarly, malolactic fermentation is desired in some white wines. FW finished wine and always involves clarification and stabilization, as in 8, 11, 12, 13, 14, 15, 33, 34, followed by 39, 41, 42. It may or may not include maturation (38) or bottle age (40), as indicated for usual styles. Stillage and lees may be treated to recover potassium bitartrate as a by-product. Pomace may also yield red pigment, seed oil, seed tannin, and wine spirits as by-products. Sweet wines are the result of either arresting fermentation at an incomplete stage (by fortification, refrigeration, or other means of yeast inactivation) or addition of juice or concentrate.

Fig. 1. Operations or actions:

1. analyze

2. select

3. harvest

4. carbonic maceration

5. destem

6. crush

7. pectinase treat

8. SO₂ addition

9. drain

10. press

11. settle

12. rack

13. centrifuge

14. filter
15. microfilter (microbially stabilize)

16. inoculate (yeast, lactic bacteria)

17. alcoholic fermentation

18. malolactic fermentation

19. acetification

20. protect from air

21. aerate

22. refrigerate

23. temperature control (generally low)

24. heat (pasteurize)

25. distill

26. concentrate (usually vacuum distill)

27. alcohol removal (reverse osmosis, vacuum distill)

28. cap management (punch down, pump over, irrigate)
29. pomace contact

30. fortify (wine spirits addition)

31. champagnization, CO₂ retention

32. carbonation

33. fine, protein stabilize

34. chemically stabilize (refrigerate, ion exchange)

35. dehydrate

36. sweeten

37. blend

38. bulk maturation (barrel, tank)

39. bottle

40. bottle age

41. case

42. market

Table 7. California Wine Shipments to All Markets^a

Year	Total, 10 L	Wine type or parameter									
		Relative amount ^b	Table, °	Sparkling, °	Dessert, °	Vermouth, °	SNW	>14, ^b °	SNW	<14, ^b °	Coolers, °
1985	1453.8	93	65.6	6.8	4.3	0.9	2.9		3.1		15.3
1986	1571.6	100	61.2	6.4	3.7	0.8	2.8		2.7		20.6
1987	1559.3	99	63.0	6.0	3.9	0.8	3.1		2.3		20.0
1988	1505.7	96	67.0	6.2	3.8	0.8	3.2		2.0		16.8
1989	1441.6	92	68.2	6.0	3.4	0.9	3.3		2.9		15.3
1990	1410.1	90	69.2	5.9	3.4	0.8	2.5		2.8		15.4
1991	1324.6	84	71.6	6.2	3.0	0.7	2.7		4.1		11.6

1									
199	1360.0	87	75.0	6.1	2.9	0.8	2.4	4.5	8.3
2									
199	1288.1	82	78.6	6.4	2.9	0.8	2.0	4.3	5.0
3									
199	1298.0	83	81.1	5.8	2.9	0.7	1.6	4.2	3.6 ^c
4									

^a Ref. 6.
^b 1986 = 100.
^c SNW = Special Natural Wines, flavored with natural flavorings; >< 14% alcohol.

Variety Selection, Fruit Production, and Harvest. The wine grape is *Vitis vinifera* and there are estimated to be a few hundred varieties of some note commercially worldwide. The list of important, widely planted varieties is much smaller. The grape variety involved is one of the most important factors in the final wine's characteristics. Data on grapes processed through wineries for California, which are essentially the same as the U.S. as a whole, are given in Table 8. For the representative years shown, the fraction of the total crush attributable to wine-grape varieties rose greatly, as the fraction attributable to raisin and table-grape varieties fell. These latter varieties, except for the small proportion of muscats, are lacking in distinctive flavor and are made into white dessert (and appetizer) wines, generic white table wines, distilled for beverage and fortifying brandy, and used for juice and concentrate production. The great decrease in winery use of these varieties occurred despite a large increase in consumption of white table wines. This decrease reflects both the greatly decreased interest in appetizer dessert wines (less base wine and less fortifying spirits needed) and the shift to varietally named premium wines from the less expensive generic wines (red or white table wine, California burgundy, etc). In California, 1993 was the first year total varietal wine shipments exceeded those of generally named wines.

In much of the period shown, white wine varieties gained in share of the total, whereas red wine varieties did not. Trends in production in response to changing consumer interests are also seen in the usage of individual winegrape varieties as a percentage of all wine varieties crushed (Table 8). Palomino, desirable for sherry, dropped sixfold. Similarly Carignane and Grenache, often used in port-type wines, dropped among the reds. Also the portion contributed by miscellaneous white and red wine varieties fell as attention shifted to the better varieties for table wine. Among the white, Colombard and Chenin blanc peaked early and then gave way a bit for Chardonnay and Sauvignon blanc. Among the red wine varieties, Cabemet Sauvignon, Zinfandel, and Pinot noir have increased at the expense of others less noted for premium wines.

Table 8. Grapes Processed in California Wineries by Type and Variety^a

Parameter	Year				
	1994	1990	1985	1980	1975
varietal type, %					
raisin	9.4	10.3	19.6	26.8	33.6
table grape	4.8	6.6	7.6	6.8	9.4
wine varieties,	85.8	83.0	72.8	66.4	57.0
red wine varieties,	36.9	31.2	30.3	42.8	38.7
white wine varieties,	48.9	51.8	42.5	23.6	18.3
Total, 10 ⁶ t	2.303	2.337	2.577	2.627	1.968
wine varieties, %, red					
Barbera	3.7	4.3	6.1	8.6	6.9
Cabernet Sauvignon	7.9	4.4	3.5	3.3	3.1
Carignane	3.3	3.9	5.8	9.6	15.8
Grenache	5.4	5.3	6.1	8.8	10.0
Pinot noir	1.5	1.5	1.5	1.3	1.1
Ruby Cabernet	1.9	2.5	3.7	7.5	6.3
Zinfandel	10.3	8.7	6.0	6.0	5.8
other red	9.0	7.0	8.9	19.4	18.9
wine varieties, %, white					
Chardonnay	14.1	7.9	4.0	1.6	0.8
Chenin blanc	9.6	12.9	14.4	8.6	6.4
Colombard	24.8	32.1	28.3	14.8	14.1
Palomino	0.4	0.4	0.7	1.7	2.6
Sauvignon blanc	3.2	2.8	2.7	0.9	0.4
other white	4.9	6.2	8.3	7.9	7.8

^a Ref. 6 and compiled.

Note most of these varieties make desirably distinctive wines, and the few that usually do not, eg, Carignane and Colombard, have special characteristics such as high yield and relative quality under difficult circumstances. With most vinifera varieties the flavor is subtle and more intensity is often sought. With varieties and hybrids from other species, such as Concord from *Vitis labrusca* or *labruscana*, the flavor may be too strong and soon satiating with a meal.

Varietal labeling is an important quality factor in the United States, and indirectly elsewhere because only certain specific varieties are planted in each prestigious foreign area. U.S. law currently requires that 75% of the wine must come from the *V. vinifera* variety named on the label. Concord-type varieties only require 51%, owing to their intense, distinctive flavor. If more than one variety is named, the relative amounts must total 100%.

Vineyard site is important to wine quality and character and interacts with variety. The general climate must not be too cold, too hot, or too humid. A mild, dry climate that still induces a dormant season, like the Mediterranean area and California, is desirable. A relatively constant weather pattern year-to-year is also sought. The nearer to the limits of cold tolerance, for example, that the climate comes, the more likely are disastrous vintages. The modifying influence of close bodies of water, sun-facing slopes, or frost-resisting air drainage can make one vineyard more desirable than another nearby.

Among satisfactory vineyards, the warmer site or vintage ripens the fruit earlier with more sugar, less acid, and less flavonoids including anthocyanins. If the situation is cool, the fruit may be held on the vine as long as possible to get adequate sugar with as much decrease in acid as possible. In a warm vineyard or vintage, sugar comes up early and fruit should be picked before acidity has declined unduly.

Proper management of the vines and vineyard are important for yield of high quality wine grapes. Pruning of the dormant vine controls the next year's fruit set because cluster primordia for next year's crop are formed during the current year. Pruning canes too long will allow more fruit to start than the photosynthetic capacity of the vine can ripen. Such overcropped fruit develops sufficient sugar for good wine either late or never and has other deficiencies, such as greater loss of acid in hot vineyards owing to delayed harvesting.

There are many details of good vineyard management as there are for any crop, but grapes are relatively tolerant, being perennials with deep roots and not high in water or fertility requirements. A few pests are special, notably grape phylloxera, a root louse native to America, but spread nearly

worldwide. It devastates *Vitis vinifera* and requires that the vine be grafted to a resistant rootstock derived from certain native vines of other *Vitis* species.

Having made the most of the available vineyard, chosen the variety and rootstock, and properly managed the vineyard, the winegrower must watch the ripening fruit and determine the optimum stage for harvest. Varieties differ in ripening date and one variety in the same vineyard may ripen at considerably different times in different years. The proper stage of grape ripeness is different for different wines. Sparkling wines need lower sugar and more tartness; dessert wines need more sugar but need to avoid raisining. Acid drops as sugar accumulates, more so in warmer places. At a minimum, the decision to harvest is based on sugar content, acidity, and close observation of the fruit's manner of changing in other parameters, including taste.

Depending on a number of variables, sugar in the range of about 20–25 Brix is chosen, with sparkling wines at the low end and dessert wines at the high end of this range. Brix is calibrated using solutions of sucrose by weight in water, ie, 1 Brix = 1 g sucrose / 100 g of solution. It is measured by hydrometry or refractometry of a juice sample and includes dissolved solids other than sugar, although sugars are on the order of 90% of the value. Higher than about 25 Brix indicates shriveling or raisining (desirable for a few wines); lower than about 20 Brix indicates underripe or overcropped grapes.

Acid content calculated as tartaric acid is about 6–7 g / L for best flavor and stability. It is higher for tart low Brix musts and less important for sweet high Brix musts. High acid levels coincide with a higher level of the second acid of grapes, malic acid.

Must Processing. Ideally the grapes are picked quickly at the chosen stage and transported to the winery so as to not cause unprotected damage to them. Mechanical harvesting is suitable, given adequate prevention of oxidation, premature fermentation, etc. As soon as possible after harvest, the grapes are ordinarily passed through destemmer–crusher machinery to remove the stems and break each berry open. If white table wine is to be made, rapid juice separation is recommended, whereas for red, fermentation of the whole pomace (pulp, skins, seeds) is usual. The method and extent of juice (or later wine) recovery from the crushed grape is important. Free-running, readily separated fluid gives the lighter products; heavier pressing leads to heavier and eventually coarser wines. There are several types and capacities of equipment for destemming, crushing, pumping, pressing, juice clarification, etc (23). Much of it is specially designed for winery use, with diverse types and capacities for special circumstances. General principles common to other food-processing equipment are considered with special care to avoid generating fine dispersal of solids, unwanted incorporation of air, or contribution of contaminants (including iron and copper) to the wines.

Because all the red color is in the epidemis of most wine grapes, proper conditions can give a white or pale pink wine from red grapes. This is the source of the recently popular "white" Zinfandel wine. The must, ie, the mixture to be fermented, is often immediately clarified for white wine by settling or centrifugation. Pectinase enzymes may be added to produce water-clear juice and better juice yield. The must comprises the whole crushed mass (pomace, ie, seeds and skins, plus juice) of red grapes, if the wine is to be red. Frequently, but not invariably, a small amount, 60 mg / L or so, of sulfur dioxide in one form or another is added to minimize any effects of oxidation and inhibit any undesirable microorganisms. If it is added, it should be as or before the grapes are crushed, because an additional effect is to inhibit polyphenol oxidase and browning.

Fermentation. Today (ca 1997) it is almost universal to inoculate the must with a selected yeast strain. Yeasts are chosen for conducting predictable, prompt, and complete fermentations under the conditions applicable for the particular wine. It is true, at least in most wineries, that grapes will ferment with the yeasts naturally present. At one time it was argued that part of the special regional character of wines was the result of the local yeasts. This is seldom claimed today, and inoculation at about 5×10^6 cells / mL is usual, using a selected, commercially grown strain, usually in the active, dry form.

Among the reasons for choosing a particular yeast strain are its ability to ferment at cold temperatures or under pressure, separate for clarification easily, or avoid production of hydrogen sulfide during typical fermentations. Wine yeasts (different strains of *Saccharomyces cerevisiae*) ordinarily do not produce characteristically different flavors. Certain "wild" yeasts can and, although they are inhibited at perhaps 6% alcohol, may produce uncontrollable or off flavors persisting from the early stages of fermentation. They can be inhibited by the addition of low levels of sulfur dioxide, because the wine yeast is relatively resistant. One of the reasons that heavy inoculation (high cell numbers) with wine strains is now usual is to raise the alcohol level quickly and overwhelm the wild forms without needing as much sulfur dioxide to control them.

The kinetics of and factors affecting fermentation are now well studied (23). Heat is released as well as a considerable volume of carbon dioxide (2 moles per mole of sugar fermented, about 50 times the must volume). To avoid excess pressure in closed fermentors, provision for escape of the CO₂ without admission of oxygen from the air is required. Uncontrolled, the temperature would rise about 1.37°C per Brix fermented. Since the Brix by hydrometer (refractometers are affected by ethanol's refractive index) falls through zero during fermentation, the original must Brix can be used for this estimate. Similarly, the ultimate alcohol content at dryness (all fermentable sugar consumed) is about 0.55 times the must Brix or 11.0–13.7% by volume for the 20–25 Brix range exemplified. This is nearly theoretical (2 C₂H₅OH per C H₁₂O), allowing for the conversion of units, etc in both cases. Note that it is reasonable to place wine tax and classification demarcations beginning at 14% alcohol.

Heat liberated during fermentation must be removed to prevent ill effects. In small fermentors such as barrels this can be by loss to ambient conditions, but in larger ones the temperature rise can become inhibitory, even fatal at about 35°C, to the wine yeast. Such a fermentation stopping before proper completion is said to be stuck and is difficult to restart. Sticking from this cause is prevented by circulating refrigerant through jackets on stainless steel tanks to control the fermentation temperature at the desired level. A large winery needs large refrigeration capacity. White table and sparkling wines are ordinarily better (more fruity, etc) if fermented cool, at about 10–16°C. Of course, the fermentation is slower the lower the temperature.

Red table wines should be fermented at about 21–29°C for proper character. They are fermented with the pomace, and carbon dioxide generation buoys the grape skins up to form a cap. For proper extraction and to prevent excessive temperatures in this cap, it is mixed with the main volume at least twice a day by punching it down or pumping the wine over it. Other procedures involve fermentors with internal baffles to submerge the cap. In modern, closed fermentors, regularly scheduled pumping from lower levels through a sprinkler over the cap is usual. This steeping and fermentation on red pomace is varied from a few hours for pink wines, down to about 6 Brix (1–5 days) for maximum color, and up to several weeks for more complex wines to be aged for the premium market.

In addition to alcoholic fermentation, a malolactic fermentation by certain desirable strains of lactic acid bacteria needs to be considered. Occasionally, wild strains produce off-flavors. Malolactic fermentation is desirable in many red table wines for increased stability, more complex flavor, and sometimes for decreased acidity. Selected strains are often added toward the end of alcoholic fermentation. All the malic acid present is converted into lactic acid, with the resultant decrease of acidity and liberation of carbon dioxide. Obviously this has more effect on the acidity the more malic acid is present, and this is the case in wine from underripe, too-tart grapes. Once malolactic fermentation has occurred, it does not recur unless another susceptible wine is blended.

Although it is sometimes encouraged in white wines, particularly barrel-fermented Chardonnay, this fermentation tends to lower fruitiness and be considered undesirable in other white wines unless acidity is too high. This is also true for pink and light red wines. If it occurs after bottling, a gassy, cloudy wine results. In such wines, it can be avoided by careful attention to clarification or filtration sufficient to remove the bacteria, by adding SO₂ at appropriate intervals as an inhibitor, or by pasteurization.

Sparkling Wines. A limited number of varieties, eg, Chardonnay and Pinot noir in Champagne, and Colombard and Chenin blanc in California, are primary in making sparkling wines. Direct pressing of whole clusters may be used to give the lightest, clearest initial must possible. A light, dry, moderate alcohol, tart, stable, white or nearly so table wine is the preferred base wine. Sugar is added in the precise amount needed to give the desired pressure (about 24 g / L for 6.078 × 10⁵ Pa (6 atm) at the end of refermentation). If fermented in large pressure fermentors, the refermented (now sparkling) wine is filtered and bottled under pressure. If it is fermented in individual bottles, they are subjected to a clarification process called riddling that leaves the sediment-free wine to be recorked in the same bottle.

If the wine is held on the yeast-containing lees for a year or more without off-flavor development, a special and desirable yeast-autolysis flavor develops. After clarification and before resealing the bottles, the wine is ordinarily sweetened slightly, even for the driest of wines, and more if the label so indicates. This is the only Californian grape must or wine that can be sweetened with sugar from other sources. Other countries and localities frequently allow adding sugar, if the grapes are not sufficiently ripe. In California, *V. vinifera* musts and wines other than sparkling may be sweetened by the addition of grape juice or its concentrate, but not by sucrose or other "artificial" sugar.

Wines can be made effervescent by carbonation rather than refermentation in a closed system, but that must be stated on the label. Even though

such wines have a slight tax advantage (see Regulations), they have not gained any following. Because they compete at the low end of the price range, the incentive to make them from superior wines has been lacking.

Sweet Wines. Slightly sweet table wines can be made by adding a small amount of reserved grape juice or concentrate, but they are often made today by refrigerating, centrifuging, or filtering out the yeasts before fermentation is complete. In any case, such wines are unstable and the residual sugar will ferment if not prevented from doing so. In bottle, gassy and cloudy wines result. Sulfur dioxide is not effective against wine yeasts, but helps inhibit malolactic bacteria. Sorbic acid added at about 150 mg L can control yeasts, at least long enough for wines marketed quickly. Properly used, dimethyldicarbonate (DMDC, at <200 mg L) can kill yeasts without undesirable residue. Heat can kill the microorganisms, but flavor change is then a problem. The most satisfactory procedure is to sterilize the wine by very close filtration, but this option requires a high level of expertise to execute. All equipment must be sterilized so that the wine does not encounter any source of viable cells downstream from the filter, until it is finally enclosed in a sterile bottle. Steam, sterilant solutions such as very high ethanol concentrations, etc may be used.

Table wines of higher sweetness can be made by extension of the same treatments, but Sauternes (French), Trockenbeerenauslesen (German), and specialty wines elsewhere are made from shriveled grapes. Such grapes have been colonized by the mold *Botrytis cinerea*, which makes the skin porous. In dry weather the grapes then lose water and the must becomes concentrated. Not only are special, concentrated flavors produced, but also the sugar becomes so high the yeasts cannot ferment it all and the wine remains sweet. The mold metabolizes some of the acid, so that is not concentrated equivalently to the sugar. Raisin flavor is to be avoided in such wines, as it is in the few wines made from partially shade-dried grapes, but it is featured in a few wines.

A third method of making sweet wines is by arresting the fermentation through use of alcohol so that the level reached prevents further fermentation and leaves the wine sweet. Most of the traditional sweet wines of the world, port and other types, are made by such fortification. Apparently owing to osmotic effect, both high sugar and high alcohol inhibit fermentation and more of one can substitute for less of the other. If a nearly sugar-free fermentation is coaxed along by small additions of sugar, alcohol can reach about 18° º by volume. Fortifications are made to this minimum level, as a rule. In the United States, the alcohol used must come from the type of fruit making up the wine, ie, in the context being discussed, grape-distilled spirits of high purity, but this is not always the case elsewhere.

Port-type red dessert wines require skin contact time to extract the anthocyanins, but the fermentation must be short to retain the sugar level near the 6–10° º level desired. The winemaker cannot always achieve desired composition in individual lots. In order to reach the desired standard, it is necessary to make new lots to enable blending to that standard. The right volume of a redder, less sweet wine will need to be made to bring to standard a lot with low color and more sugar, for example, while keeping the alcohol also within the desired limits.

Oxidized Wines. Oxidized color and flavor are defects in most wines, but certain ones capitalize upon oxidation. Sherries fall in this group and include wines that are at least moderately high in alcohol and may or may not be sweet. Access to air is required during processing or maturation and may take the form of chemical autoxidation (as in oloroso sherries) or aerobic microbiological metabolism. Aerobically, yeasts produce aldehydes from ethanol and other alcohols in the production of flor sherry types (fino, Manzanilla, etc). This, accompanied by more or less aging in wooden barrels, yeast autolysis, and involved blending results in a complex flavor. Some fino types are relatively light (about 15° º) in alcohol, but most of these wines are fortified. Flor sherries are first fermented dry and to about 15° º alcohol. This level is high enough to inhibit acetic acid bacteria (also aerobic) and yet allow aerobic growth of yeast, either as a surface film on the wine or submerged in aerated wine.

Other wines falling in this class are amber and usually sweet (Madeira-types, Malaga, Marsala, etc). Their oxidation also results from exposure to air and often long maturation, but may also involve heating, cooking must, or raisining to produce the particular type. Considering the wide variety of interesting oxidized and sweet wines made, it is unfortunate that interest in them has fallen to such a low level currently. The first year post-Prohibition in California that the shipment of fortified dessert wines did not exceed that of the table wines was 1969, and it has fallen steadily since then.

Post-Fermentation Processing. When the wine has been fermented and matured to the desired stage, it is clarified and stabilized so that it will remain clear and not be undesirably changed when bottled and marketed. Processing is held to the essential minimum to avoid flavor change and loss. Microbiological stabilization has already been mentioned, and growth of microbes in bottles is invariably undesirable in modern wines. At the end of fermentation and at each transfer thereafter wines are "racked," ie, the relatively clear portion is transferred without disturbing the yeasts and the other sediments in the lees. Much of the early clarification is accomplished by this method.

Chemical stabilization is considered necessary in the United States to prevent formation of crystals of potassium bitartrate in the bottled wine. This is commonly done by cooling the wine to barely above its freezing point, holding it to allow crystallization, and filtering it cold. Contact with preformed seed crystals and tests to verify that stability has been reached are helpful. Ion-exchange can be used, depending on the regulations of the country, but if hydrogen ions are exchanged, the acidity is raised. If sodium ions are exchanged for the excess potassium, the nutritional advantage of wine as a low sodium, high potassium food can be lost. Sometimes acidity is adjusted downward by the use of such agents as potassium bicarbonate or calcium carbonate. This procedure also affects stability, and calcium tartrate is relatively insoluble, but slow to precipitate from wine at very low levels. Sometimes in Europe tartrate crystals are tolerated or even welcomed as evidence of genuine wine.

In addition to microbes and tartrate crystals, precipitates and hazes can be formed in wines by proteins, phenolic substances, pectins and related carbohydrates, and combinations thereof. Finer and finer particles are removed by coarse or rough through-polishing and sterilizing filtration. A wide range of types and capacities of filters have been used in the wine industry, often incorporating a precoat or mixed feed, using diatomaceous earth as a filter acid. Finer particles often quickly plug and stop filters, so that a sequence of decreasing porosity filters is often needed.

Incipient hazes may not be removed at all by simple filtration. An array of fining procedures have been developed to achieve stable clarity in such cases. Fining agents are substances that are or become insoluble in wines, and, as they precipitate, adsorb or coprecipitation incipient sources of cloudiness. Properly used, the fining agents themselves are not retained in the wines and their effect is subtractive rather than additive.

Minerals, particularly Bentonite, are used to remove proteins that tend to cause haze in white wines. The natural tannin of red wines usually removes unstable proteins from them. Excess tannin and related phenols can be removed and haze from them prevented by addition of proteins or adsorbents such as polyvinylpyrrolidone. Addition of protein such as gelatin along with tannic acid can even be used to remove other proteins from white wines. Egg whites or albumen are often used to fine red wines. Casein can be used for either process, because it becomes insoluble in acidic solutions like wines.

Because each wine is likely to differ, the treatment is chosen after experimental fining tests to determine the best agent and the minimum satisfactory level. Sediment (lees) volume is held to the minimum level possible in order to avoid extra loss of wine. Usually a single well-chosen treatment is sufficient, and some wines are not fined at all. Proprietary fining agents are usually mixtures and include other agents, such as carbohydrate polymers from seaweed. These and other specialty chemicals sold for wine treatment are not large in tonnage, but are costly because of their special preparation for wine and food use.

Waste Disposal. Table-wine wineries and other large-volume food processors generate proportionally little waste. It can create a nuisance if allowed to develop off-odors, insect attraction, or oxygen demand in runoff, but generally it is innocuous. Wastes consist of stems, pomace, lees, and stillage from the production of beverage or fortifying brandy. In many instances, potential problems are avoided by frequent removal to and scattering in the vineyards. In others, stillage and wash water particularly, a chain of sealed-bottom ponds with mechanical aeration purify the water. Use in animal feeds of pomace and yeast lees is sometimes practiced. By-product recovery, whether of tartrates, distillate from fermented pomace, red pigment for food coloring, grape-seed oil, or antioxidant flavonoids and tannins, is occasionally warranted.

Storage (Maturation, Aging) and Blending. One of the prime requisites of an interesting premium wine is complexity of flavor. Maturation (bulk storage and associated final processing), aging (properly speaking, the storage of packaged wine ready for the consumer), and blending are the principal ways of achieving that complexity. Ideally, the varietal and desirable vinous characters of the wine remain prominent and recognizable, but are supplemented by these grace notes.

Maturation is conducted in closed, full containers to prevent oxidation and aerobic growth of microorganisms. Free air contact with low alcohol wine soon leads to vinegar. Except for those sherry types already mentioned, wines are exposed to air minimally and temporarily. During transfers incident to bulk storage and processing, some air exposure is almost inevitable, more in total the longer the wine is held. In the cases of white and pink table wines, it is ordinarily as near zero as possible, and stainless steel or other impermeable containers, inert gas headspace, etc are employed. Red wines withstand and even benefit from small but repeated exposures to air.

In oaken barrels, slow evaporation of water and alcohol through the staves is ordinarily compensated for by refilling and topping each barrel by

addition of the required amount of the same wine every week or so. This special process and the transfers necessary in the course of normal processing inevitably slowly permit a little oxidation, but wet wood and cork essentially do not pass oxygen (23). Oak barrels, unless exhausted by previous extraction, contribute extractives and flavor to wines. Such flavors, unless they overwhelm the other desirable characteristics of the wine in question, contribute desirably to complexity and quality.

Maturation regimes vary from as little change as possible in many white and pink wines (stainless steel tanks, cool storage, minimum time) to considerable modification in red table and a few white table wines. Fermentation and storage in fairly new 200-L barrels for about 6 mo is not uncommon for Chardonnay and white Burgundy wines. Many robust red table wines such as those from Cabernet Sauvignon grapes are often stored similarly, after fermentation and initial clarification, for up to about 3 yr in such barrels.

When they have suitably matured and been fully processed and blended, wines are bottled. Ideally, all the changes which should result during the bulk stage have been completed and the bottling itself should preserve all the desirable characteristics. Freshly bottled, the wine must be aseptic if not sterile, with minimum headspace preferably filled with nitrogen or other inert gas, and well sealed. Changes which are too slow to have been completed continue after bottling, eg, formation of ethyl tartrate, and new reactions dependent on the lowered oxidation–reduction potential can occur. Development in bottle, ie, aging proper as opposed to bulk maturation, is limited by economics to the essential few weeks or months at the winery. It may be continued in merchant and customer cellars for much longer. Special bouquets develop over several years and are the source of endless admiration and debate by the cognoscenti. The specific chemistry has not been completely clarified, but after about four years in bottle a "sun-dried linens" bouquet can be noted in many white table wines and, usually longer, a "fruity–cedar–lacquer" one in certain old red table wines. Both can be very attractive even to the uninitiated, despite some inevitable loss in grape aroma.

Conditions for proper storage of bottled wine include a fairly low and constant (about 13°C) temperature, restricted light and agitation, and bottles stored with corks wetted by the contents. Long aging of red wines often leads to some phenol-derived precipitate which can be avoided by decanting before service. Not every wine improves to the same degree, and complete bouquet development is likely to take 4–8 years for white wines and longer for red.

Regulations

Wine, along with other alcoholic beverages, was nationally prohibited in the United States by Constitutional Amendment during the period 1919 to 1933. Wine did get a unique dispensation to allow production at home for family use of 200 gallons (757 L) per year. This appears to have been in recognition of traditions of wine consumption on the part of many recent immigrants to the United States during that period. Also, there was a very small amount of commercial production for sacramental or medicinal use, but winemaking outside the home was essentially defunct. That it has been so quickly reborn and now thrives is a great tribute to the resiliency of American winemakers and their willingness to innovate.

When federal prohibition was repealed, states and localities were allowed to retain local prohibition. Today most of the states have similar laws to the federal ones, except for the imposition of additional taxes. A state can be more restrictive, but not more lenient, because then federal law would supervene. Although universal prohibition has long since been abolished, alcoholic beverages remain subject to very stringent controls and a high level of taxation, compared to other food industries. Their production is described as a permissive industry, ie, unless specifically permitted, all practices are prohibited. Some attitudes in the United States are still to an extent, holdovers from the Prohibition era; for example, a U.S. citizen can marry or become a soldier at age 18, but cannot legally consume alcohol until age 21. In the European Union, in contrast, 16 is generally the minimum legal age for drinking alcoholic beverages.

A special agency now called the Bureau of Alcohol, Tobacco, and Firearms (BATF) within the U.S. Treasury Department was empowered to regulate the alcoholic beverage industries. Although less adversarial, but strictly enforced even today (ca 1997), the regulations and their application remain voluminous and detailed. They specify not only label compliance and matters relating to taxation that are of direct interest to consumers, but contain all the details of permitted processes for and additions to wines.

The regulations have several underlying purposes: to hold composition within specified boundaries, to allow good commercial winemaking practices, to assure identity and authenticity of the product, to prevent unfair or fraudulent practices for the benefit of both the consumer and the conscientious producer, to ensure healthfulness, and to produce tax revenue. BATF regulations are not identical to those in other countries, but the United States has generally allowed import of wines made legally in other countries as long as they are properly labeled and healthfulness has not been compromised. The reverse is often not the case.

The U.S. regulations are, in effect, being continually considered for change, but the requirements for holding public hearings, the likelihood of predictable opposition, etc, make changes slow. The most recent sweeping change in the law occurred in 1990, when the entire wine code was revised and updated (25). Amendments have been (26,27) and will continue to be made, as industry practices and the prevailing legal climate evolve and change over time.

Designation was allowed for specific vineyard areas other than states and counties in 1978. These appellations have now grown to about 150 nationwide and include legally defined "Napa Valley," "Sonoma Valley," subdivisions thereof, etc, designations. A wine may be labeled as the product of a politically defined area (state, county), if 75° of the grapes from which it was made grew there. If they did not, it is labeled "American." If a specific approved appellation is designated, eg, Napa Valley, the minimum requirement is 85° from that area. If a specific vineyard estate is named, 95° of the wine must be from it. If the wine is labeled with a vintage date, 95° of the wine must have been fermented that year. Opportunity for blending of wines having these labels is severely restricted, but diversity among commercial wines is promoted.

Codification to protect regional names for wines was perfected in France and, by treaty, extended for their wines internationally under *appellation d'origine contr  lee* nomenclature (AOC, or just AC). In general these regulations delimit the region protected, limit the varieties of grapes that can be used, restrict the wine production per hectare, and require approved enological practices. Several other countries, notably Italy, Germany, and South Africa, have adopted somewhat similar regulations. It should be clearly noted that these regulations are aimed at guaranteeing type and authenticity, not necessarily quality, although there is a panel that disapproves a few wines most years which are outside the permitted range for the region.

Under these systems for assuring authenticity in regional names for wine, the larger region indicated is assumed to produce the normal minimum standard wine, and smaller subdivisions within the regions customarily attract more avid followers and command higher prices. To give a very simple example, Bordeaux includes any wine produced in the delimited district of Bordeaux. Delimited subdistricts of Bordeaux like the M  doc are further divided into specific properties, called there ch  teaux. The proprietor of a specific ch  teau zealously guards its reputation. If the wine of a given vintage from that property does not come up to the expected quality, it can be labeled and sold as M  doc, and if it is of an even more questionable quality, as Bordeaux. Owing to vagaries of vintage and the changing economic straits of owners, not to mention differences of style preference, quality fluctuates even if authenticity does not.

Every aspect of winemaking and marketing is regulated; to fully enumerate these regulations is well beyond the scope of this article. In the United States, winemakers must meet all requirements even before beginning production and must follow in detail the regulations of the BATF, particularly as listed in Title 27 of the *U.S. Code of Federal Regulations* (28). A permit must be obtained and the winery bonded before the making of wine for sale may be begun. The winery must be posted, equipped, and secured. Federal Inspectors must be allowed to make unannounced inspections. Of course, applicable state and local regulations must also be followed. For regulations in force within or those applying to the exporting of wine to European countries, the Office International de la Vigne et du Vin in Paris or the laws of the particular country in question should be consulted.

Compositional aspects that are regulated include a label statement within $\pm 1.5\%$ for table wine (for dessert wines, $\pm 1.0\%$) of the wine's alcohol content. For tax identity reasons, alcohol of 7–14° is required for table wines and 17–21° for dessert and appetizer wines. Federal excise tax rates are \$0.28 L for table wines, \$0.41 L for wines 14–21° alcohol, and \$0.83 L for those (rare) >21–24%. Coolers made with wine become taxable at 0.5° and are taxed as table wine above 7° alcohol. Sparkling wine is taxed at \$0.90 L and carbonated at \$0.87 L. The borderline to incur these taxes is CO₂ above 3.92 g L.

BATF permits no more acetic acid than 1.4 g L in red table and 1.2 g L in white and dessert wines, California and the European Union slightly less. California requires a minimum fixed acidity as tartaric of 4.0 g L for red table, 3.0 g L for white table, and 2.5 g L for dessert wines. California also requires a minimum extract in dry wines of 18 g L for red and 17 g L for white, but other states generally do not specify a minimum. In the United States, maximum total sulfur dioxide is 350 mg L. Far less is usually used today. European maxima are lower for dry wines and higher for sweet table wines.

Regulations specify a considerable list of additives and treatments which may be permitted under controlled limits and conditions. It is important to note that no wine receives more than a few of these treatments, and many have none. For example, most grape musts ferment readily without additions, but some extra nitrogen source for the yeasts is occasionally beneficial. If some is required, ammonium phosphate is the most commonly used.

Careful records must be kept to enable verification of compliance. Each lot of wine must be traceable back to the grapes and vineyard. Tanks must be carefully gauged and the capacities recorded on them. If the wine is to be labeled "estate bottled," not only must the wine be fermented, processed, and bottled by the state winery at their listed address, but the vineyard must also be owned or controlled by that winery. Other label terminology, subject to some further intricacies, are "produced," ie, fermented 75% or made into a different class of wine; "prepared," "vinted," or "cellared," ie, subjected to cellar processing or aging without changing the class of wine; "blended," ie, combined at the stated address, wines (probably purchased) of the same class and type; and "bottled" or "packed" by the stated winery.

BIBLIOGRAPHY

"Wine" in *ECT* 1st ed., Vol. 15, pp. 48–72, by M. A. Amerine, University of California; in *ECT* 2nd ed., Vol. 22, pp. 307–334, by M. A. Amerine, University of California; in *ECT* 3rd ed., Vol. 24, pp. 549–578, by M. A. Amerine, University of California at Davis.

1. B. P. Holly, *J. Wine Res.* **5**, 91 (1994).
2. V. L. Singleton, *Am. J. Enol. Vitic.* **43**, 344 (1992).
3. *Bull. OIV* (769–770), 971–1024 (1995).
4. Ref. 3 and earlier numbers.
5. *Wines Vines* **76**(7), 16–44 (1995).
6. Ref. 5 and earlier numbers.
7. L. H. Lesko, *King Tut's Wine Cellar*, B. C. Scribe Publications, Berkeley, Calif., 1977.
8. P. E. McGovern, S. J. Fleming, and S. H. Katz, *The Origins and Ancient History of Wine*, Gordon and Breach Publishers, Luxembourg, 1995.
9. R. Dion, *J. Wine Res.* **5**, 215 (1994).
10. P. T. H. Unwin, *Wine and the Vine, An Historical Geography of Viticulture and the Wine Trade*, Routledge, Chapman & Hall, New York, 1991.
11. D. I. Jackson, and P. B. Lombard, *Am. J. Enol. Vitic.* **44**, 409 (1993).
12. *Proceedings of a Symposium on Potential Health Effects of Components of Plant Foods and Beverages in the Diet*, Aug. 14–15, 1992, University of California at Davis, p. 106.
13. M. A. Amerine, *Adv. Food Res.* **5**, 358 (1954); **8**, 133 (1958).
14. L. Nykanen and H. Suomalainen, *Aroma of Beer, Wine and Distilled Alcoholic Beverages*, Akademie-Verlag, Berlin, 1983.
15. V. L. Singleton and P. Esau, *Phenolic Substances of Grapes and Wine, and Their Significance*, *Adv. Food Res.*, Suppl. 1, (1969).
16. H. F. Linskens and J. F. Jackson, *Wine Analysis*, Vol. 6 in *Modern Methods of Plant Analysis*, Springer Verlag, Berlin, 1988.
17. P. Schreier, *CRC Crit. Rev. Food Sci.* **12**, 59 (1979).
18. C. S. Ough and M. A. Amerine, *Methods for Analysis of Musts and Wines*, 2nd ed., John Wiley & Sons, Inc., New York, 1988.
19. M. A. Amerine and E. B. Roessler, *Wines, Their Sensory Evaluation*, rev. ed., W. H. Freeman and Co., New York, 1983.
20. M. E. Weisse, B. Eberly, and D. A. Person, *British Med. J.* **311**(7021), 1657 (1995).
21. L. F. Bisson, C. E. Butzke, and S. E. Ebeler, *Am. J. Enol. Vitic.* **46**, 449 (1995).
22. *Report of the Dietary Guidelines Advisory Committee on the Dietary Guidelines for Americans, 1995*, U.S. Department of Agriculture, Department of Health and Human Services, Washington, D.C., 1995, pp. 52.
23. R. B. Boulton, V. L. Singleton, L. F. Bisson, and R. E. Kunkee, *Principles and Practices of Winemaking*, Chapman & Hall, New York, 1995, pp. 8–9.
24. C. S. Ough, *Winemaking Basics*, Haworth Press, Binghampton, N.Y., 1992.
25. *Fed. Reg.* **55**(118), 24974–25034 (1990).
26. *Fed. Reg.* **58**(193), 52222–52232 (1993).
27. *Fed. Reg.* **60**(227), 58311–58318 (1995).
28. *U.S. Code of Federal Regulations*, Title 27, Chapter 1, Subchapter A, Parts 1, 4, 9, 12, 16, 24 (Apr. 1 1997), pp. 243.

Vernon L. Singleton
Christian E. Butzke
University of California, Davis

WINTERGREEN OIL.

See OILS, ESSENTIAL.

WITHERITE.

See BARIUM COMPOUNDS.

WOLFRAM AND WOLFRAM ALLOYS.

See TUNGSTEN AND TUNGSTEN ALLOYS.

WOOD-BASED COMPOSITES AND LAMINATES

Various trees provide many things of use to people in addition to wood and paper products, such as heating and cooking fuel, fruits, nuts, chemicals, and useful drugs. Trees provide homes for wildlife and birds as well as beautiful scenery. Perhaps most importantly, they use sunlight to convert carbon dioxide from the air and water as the basic building blocks to produce their woody substance, bark, and leaves. A by-product of these photosynthesis reactions is oxygen, which is emitted into the air. Thus, trees as well as all green plants are essential in maintaining acceptable levels of oxygen and carbon dioxide in the atmosphere.

As trees grow, cellulose molecules are produced. These molecules are long strands of 6-carbon ring structures attached end-to-end. They may have a length of as much as 1500 units. Many cellulose molecules are laid side-by-side to produce microfibrils. The microfibrils form the basic structure of the wood fibers. Most fibers develop so that their strongest, lengthwise direction is parallel to the vertical direction of the stem. A substance called lignin is formed to reinforce and bond the fibers to one another. As the tree grows, these layers of fibers develop outward and upward to form the stem of the tree. These natural, renewable structures (fibers and wood) are the basis of virtually all major uses of forest products.

As with all living things, trees have a finite lifespan. Only a few species will flourish for more than 100 years, the majority passing into old age at 30 to 80 years. Because wood, wood products, and paper make up such an important and integral part of our lives, it is equally important to have well-conceived forest management plans that provide for growth management, timely removals, and speedy regeneration. Only in this way can adequate supplies for future generations be assured. Fortunately, the United States is endowed with some of the best lands and climate for forest production in the world. Properly managed, these lands could supply the future needs not only of the United States but a significant portion of the needs of other parts of the world as well. In addition, these lands can provide significant employment and resultant contributions to the economy.

HISTORICAL DEVELOPMENT

Wood is one of the oldest construction materials in human use and continues to be an extremely valuable material to this day. Wood is lightweight, strong, stable, easily worked and fastened, a good insulator, warm to the touch, and pleasing in appearance, among many other attributes.

Logs and rough beams were probably the first uses of wood, employed in large construction projects as soon as simple tools became available. Prior to the industrial revolution, sawn lumber was available only in limited quantities. Paper, also in small quantities, was made from cotton rags, or papyrus, the inner bark of some species of trees, or rice straw in China and Japan. The advent of large machinery allowed production of lumber as a commodity. The manufacture of large quantities of wood veneer followed quickly. The discovery of chemical pulping of wood to separate the fibers and lignin coincided with the availability of large machinery to make continuous lengths of paper. Thus, the wood products and paper industries began, without which a modern way of life would have been impossible to achieve.

The vast natural resources of timber in the United States, through the foresight of early foresters and political leaders, led to the availability of large quantities of lumber at reasonable cost. This abundance has made the United States the best-housed and most comfortable of all the areas of the world. However, in the conversion of timber to lumber or veneer, only about one-half of the volume of the log actually becomes lumber or veneer. The remainder is in the form of bark, slabs, edgings, sawdust, planer shavings, veneer round-up waste and clippings, veneer cores, and plywood trim waste. For some years, these residues were largely unused except for small amounts diverted to pulp and paper mills. Then, in the early 1900s, various developers and entrepreneurs began discovering new and useful products which could be made from these residues. Today (ca 1997), virtually the entire woody stem of the tree is converted to a family of wood and wood-based products. These developments have literally doubled, and perhaps with recycling, more than doubled the yield of usable products from a given volume of wood, contributing even more greatly to the comfort of ordinary people and to the economy. The products that have resulted are treated herein in chronological order of development as much as possible. Inasmuch as most of these products are bonded together with adhesives, the major types of wood-bonding adhesives are also treated briefly.

Adhesives for Wood

The following list includes most of the adhesives which can be or have been used in wood and wood composite bonding applications.

Natural adhesives

- starch
- soy flour
- blood
- casein
- hide bone

Synthetic adhesives

- urea–formaldehyde

- melamine–formaldehyde
- phenol–formaldehyde
- resorcinol–formaldehyde
- isocyanate urethanes
- poly(vinyl acetate) (PVA)
- contact adhesives
- epoxy resins
- hot-melt adhesives

In the natural adhesive group, only limited quantities are still used today, largely in conjunction with a synthetic adhesive where the purpose of the mixture is to improve some characteristic of the adhesive. In early plywood hot-pressing applications, blood and starch or soy flour were used. Casein, a milk-based product, was a long-time favorite adhesive for cold-pressed plywood and lumber laminates. Soy flour glues were used in the early years of cold-pressed softwood plywood. Hide and bone glues constituted the original version of the hot-melt glue and were widely used as assembly glues in furniture construction. A disadvantage of the natural glues was that they were suitable only for interior uses, where exposure to liquid water would not be encountered. Casein and blood-soya glues were the most water-resistant, but neither could be classed as truly suitable for exterior exposure.

Synthetic adhesives have resulted from developments in the field of organic chemistry. Virtually all are obtained from petroleum, natural gas, or coal by-products.

Formaldehyde, HCHO, is a primary and necessary constituent of the first five synthetic adhesives in the listing. It is a simple organic chemical first identified during the latter half of the 1800s. Its irritating and toxic odor and preservative properties were known from the time of its early development. It is a ubiquitous chemical, formed naturally in small quantities by every process of incomplete combustion as well as in normal biologic processes. The human body has a natural formaldehyde level of about 3 µg g, ie, 3 parts per million (ppm) in the blood at all times.

During the late 1970s, concerns were raised about levels of airborne formaldehyde in buildings resulting primarily from construction using composite panels bonded with urea–formaldehyde resins and combined with energy-efficient building practices which reduced air losses.

In an effort to address these concerns, regulations and guidelines were established. The U.S. Department of Housing and Urban Development (HUD) regulation 24CFR 3280.208, *Manufactured Home Construction and Safety Standard*, promulgated in 1985 (1), required emissions of 0.2 µL L (ppm) or less from hardwood plywood and 0.3 µL L (ppm) or less from particleboards, when both products are made with urea–formaldehyde adhesives and ultimately used in manufactured homes. The industries implemented voluntary standards, and formaldehyde emission levels were significantly reduced. In 1993, the National Particleboard Association voluntarily reduced emissions to 0.2 µL L (ppm) or less for particleboards used in flooring applications. Airborne formaldehyde levels in most studies of conventional homes have shown average levels of less than 0.1 µL L (ppm), a level which also is generally accepted by the U.S. Environmental Protection Agency as a level of concern. It should be noted that the average person cannot detect formaldehyde in air below levels of about 0.5 µL L (ppm), although there have been reports of hypersensitive individuals. Another factor which should ease concern about formaldehyde is that most emissions occur as a result of formaldehyde trapped in the product during manufacture. As the formaldehyde is slowly released over time the natural emission rates decrease, resulting over time in decreasing levels in the surrounding air. Today there should be little concern for irritation problems from formaldehyde in most current construction and there should be no concern about cancer caused by breathing formaldehyde in normal air.

Phenol–formaldehyde (PF) was the first of the synthetic adhesives developed. By combining phenol with formaldehyde, which has exceptional cross-linking abilities with many chemicals and materials, and a small amount of sodium hydroxide, a resin was obtained. The first resins solidified as they cooled, and it was discovered that if it was ground to a powder with a small amount of additional formaldehyde and the application of more heat, the mixture would liquify and then convert to a permanently hard material. Upon combination of the powdered resin mixture with a filler material such as wood flour, the result then being placed in a mold and pressed under heat and pressure, a hard, durable, black plastic material was found to result. For many years these resulting products were called Bakelite, the trade name of the inventor. Bakelite products are still produced today, but this use accounts for only a small portion of the PF resins used.

Subsequently it was discovered that liquid PF resins could be produced and used as binders for other products. The primary uses for one class of PF resins is in foundry applications where the resin is employed in binding the sands used in the molds. The resin burns away during the molding process, allowing removal and recycling of the sand. Today the major use of PF resins is in bonded wood products, ie, plywood, hardboard, hardboard siding, oriented strandboard, and waferboard. PF resins produce strong, durable, and waterproof bonds; primary characteristics required in any product which may be exposed to the weather or to liquid water for significant time periods. PF resins generally produce a dark-colored glueline, require hot-pressing to cure, and are much more expensive than the popular urea-based adhesives used for interior exposures.

Urea–formaldehyde (UF) was also developed as an early synthetic adhesive by combining the two chemicals under carefully controlled heating (cooking) conditions. Small amounts of other chemicals are added to obtain other desired characteristics in the resin adhesives. Urea–formaldehyde resins are inexpensive, fast-bonding, produce a colorless glueline, are water-resistant, and tolerate a fairly wide range of bonding conditions. These resins are an almost ideal interior-grade adhesive, with the single exception of formaldehyde emissions. However, developments in improved resin chemistry have reduced potential emissions to unnoticeable levels, with a possible exception occurring in the case of a few hypersensitive individuals. Urea–formaldehyde resins cure at acid conditions which means that they can cure at room temperature with the simple addition of a small amount of acid catalyst, or more quickly under heat and pressure using the natural acidity of the wood as a catalyst.

Both melamine–formaldehyde (MF) and resorcinol–formaldehyde (RF) followed the earlier developments of phenol–, and urea–formaldehyde. Melamine has a more complex structure than urea and is also more expensive. Melamine-base resins require heat to cure, produce colorless gluelines, and are much more water-resistant than urea resins but still are not quite waterproof. Because of melamine's similarity to urea, it is often used in fairly small amounts with urea to produce melamine–urea–formaldehyde (MUF) resins. Thus, the improved characteristics of melamine can be combined with the economy of urea to provide an improved adhesive at a moderate increase in cost. The improvement is roughly proportional to the amount of melamine used; the range of addition may be from 5 to 35%, with 5–10% most common.

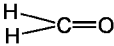
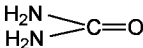
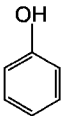
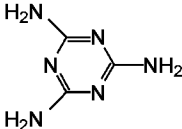
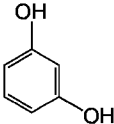
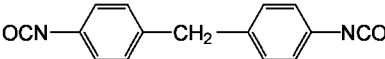
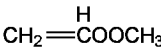
Resorcinol is to phenol as melamine is to urea. Resorcinol–formaldehyde (RF) is very expensive, produces dark and waterproof gluelines, but will cure at room temperature. As with melamine and urea, resorcinol is often combined with phenol to produce phenol–resorcinol–formaldehyde (PRF) adhesives, thus producing an excellent adhesive with some of the economy of phenol. These adhesives are the mainstay of the laminated timber industry which generally requires a room-temperature cure with durable, waterproof gluelines.

Another class of adhesives has been gaining more widespread use in the wood laminates and composites industries. These adhesives are isocyanates, the common name of a chemical group of diphenylmethane-*p,p*-diisocyanate polymers. They are excellent adhesives, fast-curing, and will produce durable, waterproof bonds if used in sufficient quantity. Considerations which have slowed wider usage of isocyanates are high cost (about 6× urea resins and 3× phenol resins); adhesive qualities which bonds to press plates as easily as to wood, thus requiring a totally reliable press release additive system; and finally, their toxic characteristics require a high level of manufacturing safeguards. The liquid adhesive should not contact the skin, and whereas only very minor amounts of odorless emissions occur during application and curing, about 4% of exposed persons are sensitive to these fumes and may develop a life-threatening asthmatic condition on continued exposure. However, their fast curing rate offsets some of the higher cost as compared to phenolics, and the other primary concerns can be handled by special care in monitoring, handling, and dust ventilation controls. Provided they are used properly and safely, isocyanates have been found to be excellent bonding agents.

Four other groups of synthetic adhesives find uses in secondary processing, ie, overlaying, assembly gluing, etc, and in furniture and cabinet manufacture. Poly(vinyl acetate) (PVA) adhesives are widely used in application of veneers and other overlays to panel substrates and in some unit-assembly operations. PVA adhesives are an emulsion of polyvinyl acetate in water and cure by loss of water. The PVA adhesives are somewhat

expensive, but are extremely versatile, cure at room temperature, and provide a colorless glue. The curing reaction time may be decreased by addition of mild heat, or by using a catalyst in the case of special types of PVA resins. Table 1 lists the chemical formulas of all of the preceding chemicals widely used in wood-bonding applications.

Table 1. Chemical Structures of Constituents of Principal Wood Adhesives

Name	Chemical formula	Chemical structure
formaldehyde	(HCHO)	
urea	(H ₂ NCONH ₂)	
phenol	(C ₆ H ₅ OH)	
melamine	(C ₃ H ₃ N ₃)	
resorcinol	(1,3-(HO) ₂ C ₆ H ₄)	
isocyanate	(NCO) ₂ CH ₂ (C ₆ H ₄) ₂	
poly(vinyl acetate)	(CH ₃ CO ₂ CH=CH ₂) _n	

Another widely used overlay adhesive is the contact type. These specialized adhesives, in the same group as rubber cement, may be of the solvent-base or water-base types. They are often used to bond overlays such as wood veneer, vinyl (poly(vinyl chloride)) films, or high pressure laminates such as countertop overlays.

Epoxy resins are also used in special applications, such as an overlaying procedure requiring a durable, heat-resistant bond of a difficult-to-bond overlay on a wood-base panel substrate. Metal sheets used as overlays, for example, often require an epoxy adhesive.

Finally, a large variety of hot-melt or thermoplastic adhesives have been developed in recent years. These are solid at normal temperatures, but melt and flow if heated and resolidify when cooled. A wide range of melting points and bond strengths are available, depending on the requirements of the application. These adhesives are widely used in furniture and cabinet construction and have largely replaced the hide bone glues formerly used in these applications.

Assuming correct adhesive application and bonding conditions, it is generally agreed that the bonding mechanisms are the result of forces of molecular attraction between the adhesive and the wood surfaces. Any normal structural wood adhesive should be capable of producing bonds which are stronger than the shear strength of the wood. Treatments of the manufacture and uses of the various composite and laminated products given herein provide more specific details on the applications and uses of adhesives.

MANUFACTURING PROCESSES

SAFETY

An extremely important safety issue with respect to all wood product manufacturing processes is personal worker safety. All of the processes use much moving machinery, usually including many saws or knives. Workers must continually remember the inherent dangers these machines involve as well as other possible dangerous situations which could result from malfunctions or other errors. In addition, most processes are more or less dusty and noisy. Most employers require use of safety glasses and many require hearing protection, safety shoes, and hardhats as well as other kinds of protection needed for specific jobs.

Plywood

Plywood is a panel made from wood veneers (thin slices or sheets) bonded to one another. Generally each ply is oriented at right angles to the adjacent ply, and the two face plies should have the grain direction parallel to each other. Thus most plywood will have an uneven number of plies, such as 3, 5, 7, or more. An exception to this is a four-ply construction in which the two core plies are oriented parallel to one another and perpendicular to the two face plies.

Plywood has essentially equal stability in both panel directions and is almost as stable as the parent wood in the direction of the wood grain. Strength properties in bending are roughly proportional in each panel direction to the amount of wood in those layers closest to the surface which are parallel to the wood grain direction. Thus as the number of plies increases, these bending properties become more equalized in both panel directions.

HARDWOOD PLYWOOD

Hardwood is generally considered to be that coming from a tree having distinct leaves as opposed to needles, and usually but not always deciduous, ie, growing and losing a set of leaves each year. Whereas examples of small wood veneers bonded to other woods have been found in ancient tombs and other historical locations, wide usage of hardwood plywood did not occur until slicing and peeling equipment became available to produce veneers in reasonably sufficient quantities.

Hardwood Plywood Processing and Products. Production of hardwood plywood follows one of two general processes. These are illustrated in Figures 1 and 2. High quality face veneers are almost always sliced from halved or quartered logs in order to obtain the maximum quantity of highly figured veneer. These veneers are extremely thin, in the 0.4–0.7-mm thickness range. Some of the principal species used for decorative veneers are walnut, maple, birch, oak, as well as many tropical species, including teak and rosewood.

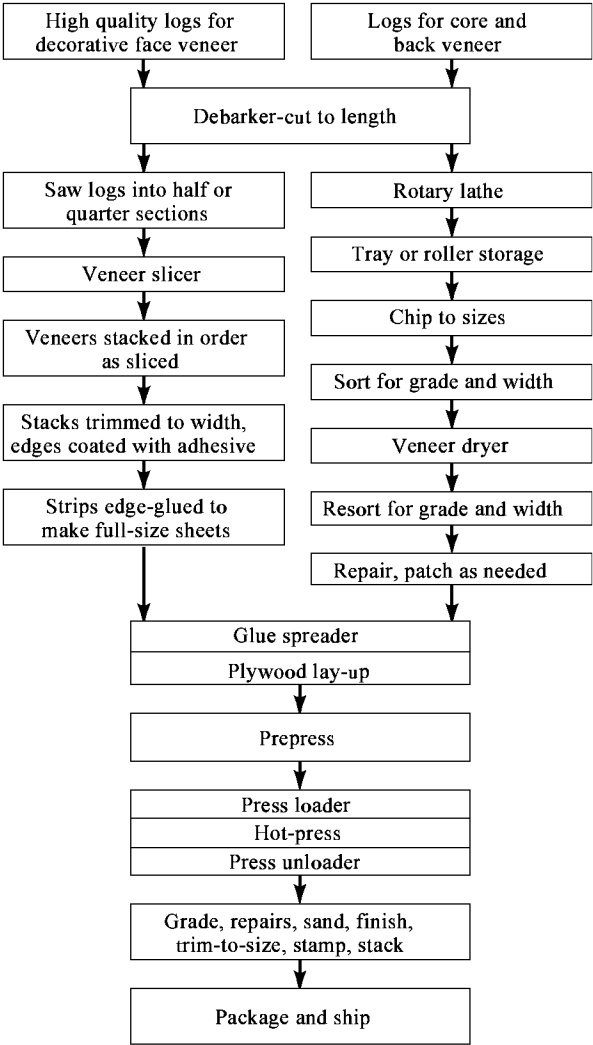


Fig. 1. The hardwood plywood process using sliced decorative veneers.

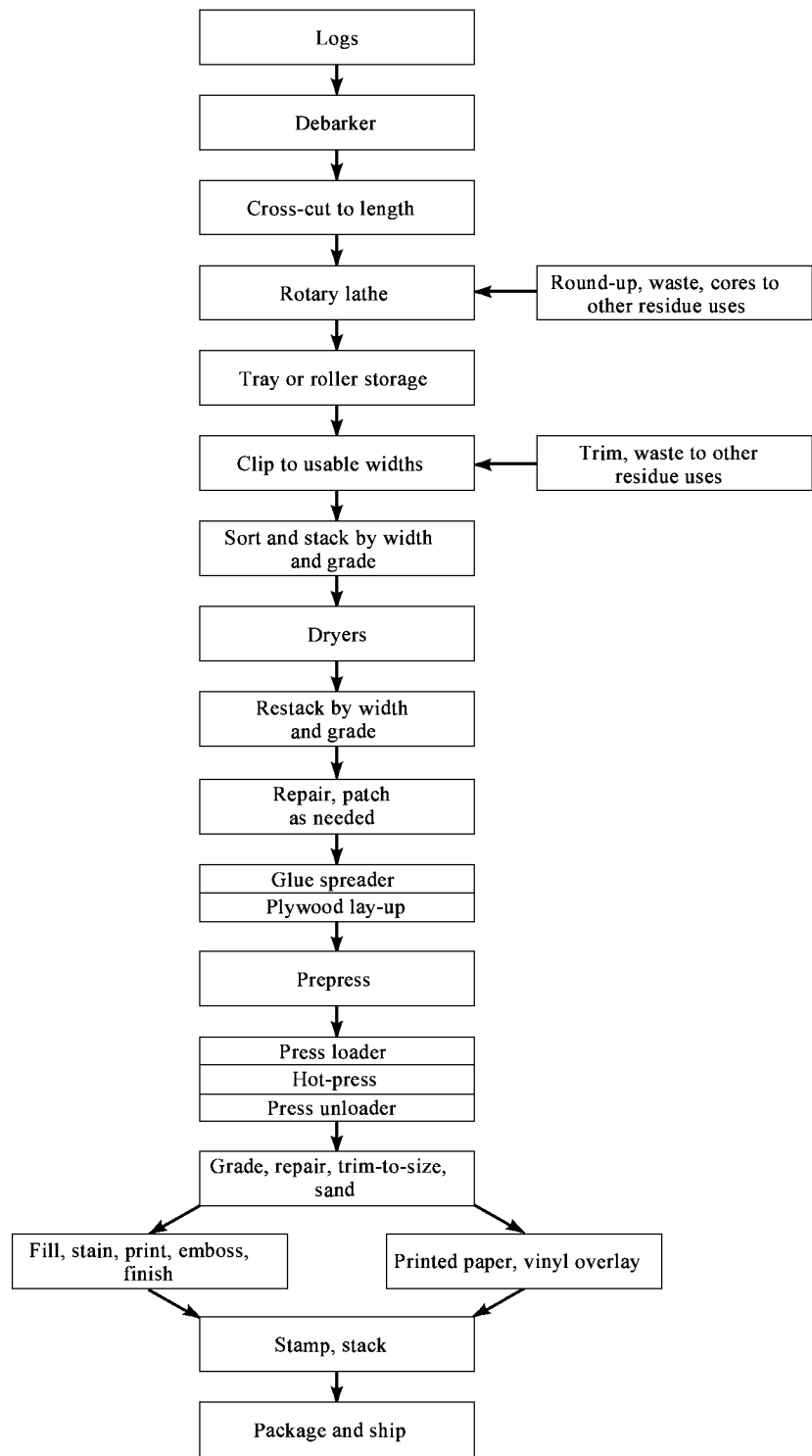


Fig. 2. The hardwood plywood process using rotary-cut veneers.

Hardwood plywood logs are almost always soaked or steamed to soften the wood, making it more amenable to slicing or peeling. If the logs are to be sliced, they are usually sawn into halves or quarter sections. These are called flitches. The slicer uses a long knife to move and slice across the flitch, producing a thin strip of veneer. The knife then retracts and the flitch is indexed forward by the desired veneer thickness and sliced again. This cycle is repeated until no more veneer can be produced from the flitch. Veneers from each flitch remain together during the entire process so that wood grain patterns can be matched in the final panels.

Other veneers for backs and cores, or in the case of the tropical luauns, are almost always peeled on a rotary lathe. In this process, the log is rotated against a long, stationary knife and the log is essentially "unrolled" in a long, thin sheet. These sheets are called rotary-cut veneers.

Decorative veneers from the slicer are carefully passed through a long tunnel dryer. As they emerge from the dryer at 3–5% moisture content (MC), they are carefully restacked in the order they emerge. They proceed to a trimmer, which trims the stack to obtain parallel edges, at the maximum obtainable width. A coating of adhesive, usually a contact adhesive, is applied to the edges of the stack. Then the pieces are passed edge-to-edge through an edge-gluer, from which sheets of increasing width result with each pass through the machine. When the desired width, usually about 1270 mm, is achieved, the sheets are carefully stacked and moved to the panel lay-up area.

Veneers from the rotary lathe pass through a high speed clipper or knife which cuts out the usable widths, the objective being to obtain as many full-width sheets as possible. These veneers move along a table where they are sorted into piles by width and appearance grade. These piles of veneer are then fed sheetwise into long tunnel dryers. The veneers should be at 3–5% MC as they emerge from the dryers. Moisture sensors are used to check each piece, moisture control being an extremely critical part of the process. Veneers which are too wet are marked and placed in stacks to be redried. Veneers are again sorted after emerging from the dryer, special attention being given to grade and to those veneers in need of repair or patching. Because these are not face veneers, patching and repairs are limited to those required to allow the veneer to reenter the desired grade. This may involve simple taping of splits and tears or plugging of an oversize knot area with a wood plug. In the case of the latter, patches or plugs of various shapes are punched from strips of veneer. Then the plugging machine punches out a section of similar shape around the knot or defect and replaces it with a veneer plug of the same shape. Common shapes are boat, dogbones, and ovals.

The stacks of veneer are moved to the panel lay-up area, where adhesive is applied and veneers assembled into lay-ups ready to be pressed. Because core cross-veneers are only half-panel width, these stacks are cut into halves.

The primary adhesive used in hardwood plywood is urea–formaldehyde (UF) mixed with wheat flour as an extender to improve spreadability, reduce penetration, and provide dry-out resistance. A catalyst may also be added to UF resins to speed the cure or to cause the UF to cure. Scavengers also may be added to reduce formaldehyde emissions from finished panels. If more water-resistance is required using a UF bond, small amounts of melamine may be added, producing a melamine–urea–formaldehyde (MUF) adhesive.

For exterior applications, where water exposure is expected, phenol–formaldehyde (PF) or phenol–resorcinol–formaldehyde (PRF) adhesives are used. Only small quantities of this type of hardwood plywood are made, primarily for marine use.

Lay-up proceeds by laying down the veneer which is to be the back surface of the panel. Then a sufficient number of pieces of core veneer are passed through the glue spreader to form the next layer of cross-oriented veneer. The glue spreader commonly used in hardwood plywood manufacture is a roll coater in which a pair of opposing rubber rolls are coated with a thin layer of adhesive. As the veneer is passed between the rolls, the adhesive is transferred to the surfaces of the veneer. Adhesive is applied only to the cross-plyies and in sufficient quantity to provide a continuous layer on both opposing faces of veneer. Thus, in the case of a three-ply panel, only the core layer is spread with adhesive and in that of a five-ply panel, the second and fourth layers both of which are cross-plyies, are spread with adhesive. Then the top surface veneer, which is normally the decorative surface, is placed on the assembly.

As succeeding panels are laid-up, a stack of panels is formed. If the panels are to be cold-pressed, an uncommon procedure in modern manufacturing, the stack will be high enough to fit into the cold-press. The stack is rolled into the press, the press is closed under hydraulic pressure and the bonding pressure, 1035–1205 (ca 150–175 psi) is maintained for the time required to form a bond. This time could vary from 30 to 120 minutes, depending on temperature and the adhesive formulation used.

If the laid-up panels are to be hot-pressed and, for example the hot-press has 20 openings or spaces in which to place veneer assemblies to be pressed, the stack would be 20 panels high. If, as in the case of thin panels, two assemblies were to be pressed in each opening, the stack would be 40 panels high. The stacks are moved to a large one-opening pre-press where they are cold-pressed as a unit for several minutes. This pressing completes the transfer of adhesive from the coated faces to the uncoated faces and also provides a small amount of bonding so that the individual panels have sufficient integrity to be moved individually into the desired press opening.

After prepressing, the stacks move to the press loader. The press loader raises the stack vertically beside the press. As the top panel reaches each opening, the panel is manually pushed into the press. When the loader reaches the top opening of the press and all panels are in the press, the press begins the pressing cycle, closing on each panel and pressing the entire group of panels at 1035–1205 kPa (ca 150–175 psi) for several minutes at a temperature of about 140°C. Time in the press is proportional to the thickness of the panel(s) in each opening. Heating greatly accelerates the cure of the adhesive. The combination of heat, pressure, and moisture also produces a small amount of densification in the veneers, resulting in a finished panel density which may be a few percent higher than that of the original wood.

After the allotted pressing heating time, the pressure is released and the press moves to the open position. By this time, the loader is again ready and as each veneer assembly is moved into its respective opening, the pressed panels are pushed out into the unloader on the opposite side of the press. As the press begins the next cycle, the unloader moves to deposit the pressload of panels into a stack.

The stacks are moved and again separated into individual panels where they pass a grading station. Those panels requiring touch-up or repair move to the repair stations. Panels then move to the trim saws where edges are trimmed to the final desired size, normally 1220 × 2440 mm (4 × 8 ft.). Panels are then touch-sanded to final thickness, and pre-finished as desired. Those panels with high quality decorative veneer faces are usually filled, stained to the desired tone, and finished with a clear finish.

Panels of the luaun type are mostly manufactured in the Far East by a similar process, but using rotary-cut veneers, and are then imported into the United States as raw panels. These panels usually begin their processing at a touch sander and then move to the desired finishing sequence where filling, coating, grain embossing, and final finishing steps occur. The panels may also receive a thin vinyl overlay or printed paper, followed by a clear finish. Many of these coated or overlaid panels are virtually indistinguishable from natural wood. After the finishing process, the panels are graded, defective panels removed, and the remainder packaged for shipment.

Uses and Treatments of Hardwood Plywood. Most early applications of hardwood plywood were those where the hardwood plywood was better adapted to the use than solid wood. One of the most important early uses was in curved or formed parts, an application particularly suited to the use of veneers which could be molded into intricate shapes during the pressing and bonding process. Then, as furniture manufacturers realized the inherently superior stability of plywood compared to solid wood, lumber-core or plywood panels began to be used for most flat-panel constructions in furniture.

Lumber core is a five-ply panel, usually about 19 mm (3 4 in.) thick, in which the bulk of the thickness, about 16 mm (5 8 in.) is edge-glued lumber. Yellow poplar and red gum are desired species for lumber core. Cross-plyies of lower value wood veneers are laid at right angles to the core grain direction, followed by two thin surface plyies of the decorative face veneer in a parallel direction to the core. This assembly is pressed and bonded to form a panel of exceptional quality, provided all steps are accomplished in a desirable manner.

Plywood furniture core panels, also about 19 mm (3 4 in.) thick, were normally made of a number of layers of relatively thick, 1.5–3.0 mm (1 16–1 8 in.) lower value wood veneers combined with thin surface plyies of the decorative veneer. These assemblies were laid-up from glued veneers and then pressed while the bonding occurred. Both lumber core and plywood core have been almost totally displaced in recent years by particleboard or medium-density fiberboard, both discussed herein. This change resulted from the increasing availability and improved finishing characteristics of composites and from decreasing supplies of core lumber or veneer of suitable quality.

One type of thick hardwood plywood still available is imported from the northern Scandinavian countries and is generally known as Finnish birch. Characteristically, these plywoods are manufactured using multiple layers of veneer of the same thickness, about 1.5 mm (1 16 in.), and bonded with a urea–formaldehyde or melamine–urea–formaldehyde adhesive.

Thin hardwood plywood in the range of 4.5–6.0 mm (3 16–1 4 in.) was normally a three-ply construction with a thin, medium-quality back ply, a thicker lower value core, and another thin, high quality decorative face veneer. These panels were used as wall paneling, door facings, or for furniture cabinet applications requiring thin panels. Currently, only relatively small quantities of these types of panels are produced in the United States. The majority of thin paneling used today is imported from the Far East and is made from various tropical species of the luaun group, sometimes known as Philippine mahogany. These panels are normally finished using one of the processes intended to create the appearance and grain pattern of a decorative veneer or other patterns.

Economic Aspects. The hardwood and decorative plywood industry has decreased in size and production significantly in the past few years. In 1994, there were an estimated 100 mills operating in the United States having a production volume of 1.135 × 10⁶ m³ (2). The dollar value of this production is extremely difficult to estimate because of the very wide range of prices for the products.

Specifications, Standards, Quality Control, and Health and Safety. The Hardwood Plywood and Veneer Association (HPVA) represents the manufacturers and associated industries in matters relating to specifications, standards, quality control, and health and safety factors. Specifications and standards for hardwood plywood products are found in the American National Standard Institute (ANSI) standard for *Hardwood and Decorative Plywood* (3). The most important product quality test procedures are bond quality tests and the formaldehyde emission test. The formaldehyde emission test is described in ASTM E1333-90 (4). Test panels are placed in a test chamber maintained at 25°C, 50% relative humidity (rh), 0.5 air change per hour, and having a test product loading rate of 0.95 m² of product surface m³ of chamber volume, ie, 0.95 m² m³. These are considered to be normal conditions of temperature, rh, air ventilation rate, and product usage within a manufactured home. Formaldehyde concentrations within the test chamber must be <0.20 µL L (ppm) to qualify the material for use in manufactured homes. At the time of this writing (ca 1997), there is no emission regulation for products used in other buildings, but the majority of products are now manufactured to pass this stringent emissions test. Another important test is the flamespread test, described in ASTM E-84 (5). This test is designed to measure the potential of a product to ignite and contribute to the spread of a flamefront in a fire.

SOFTWOOD PLYWOOD

Softwood is generally considered to be that coming from a coniferous tree, ie, an evergreen tree having needle-like or scale-like leaves. There are exceptions to the evergreen rule, however. In addition, many hardwoods also may now be used in softwood plywood as core veneers.

Softwood Plywood Processing and Products. The first softwood plywood was made in 1905 using Douglas fir veneer, soya flour adhesive, and cold-press system. Figure 3 illustrates the traditional softwood plywood process. Veneer logs are kept wet until cut to peeler block lengths, about 2500 mm (100 in.). The peeler block may be stored in cold or hot water, or placed in large, steam-heated rooms until it is removed to be used for veneer. The block is placed in a charger, which quickly loads and positions the block into the lathe after the previous block is finished. The charger electronically scans each block and determines a point at each end of the log which will be centered in the lathe and will allow the optimum yield of veneer from the block.

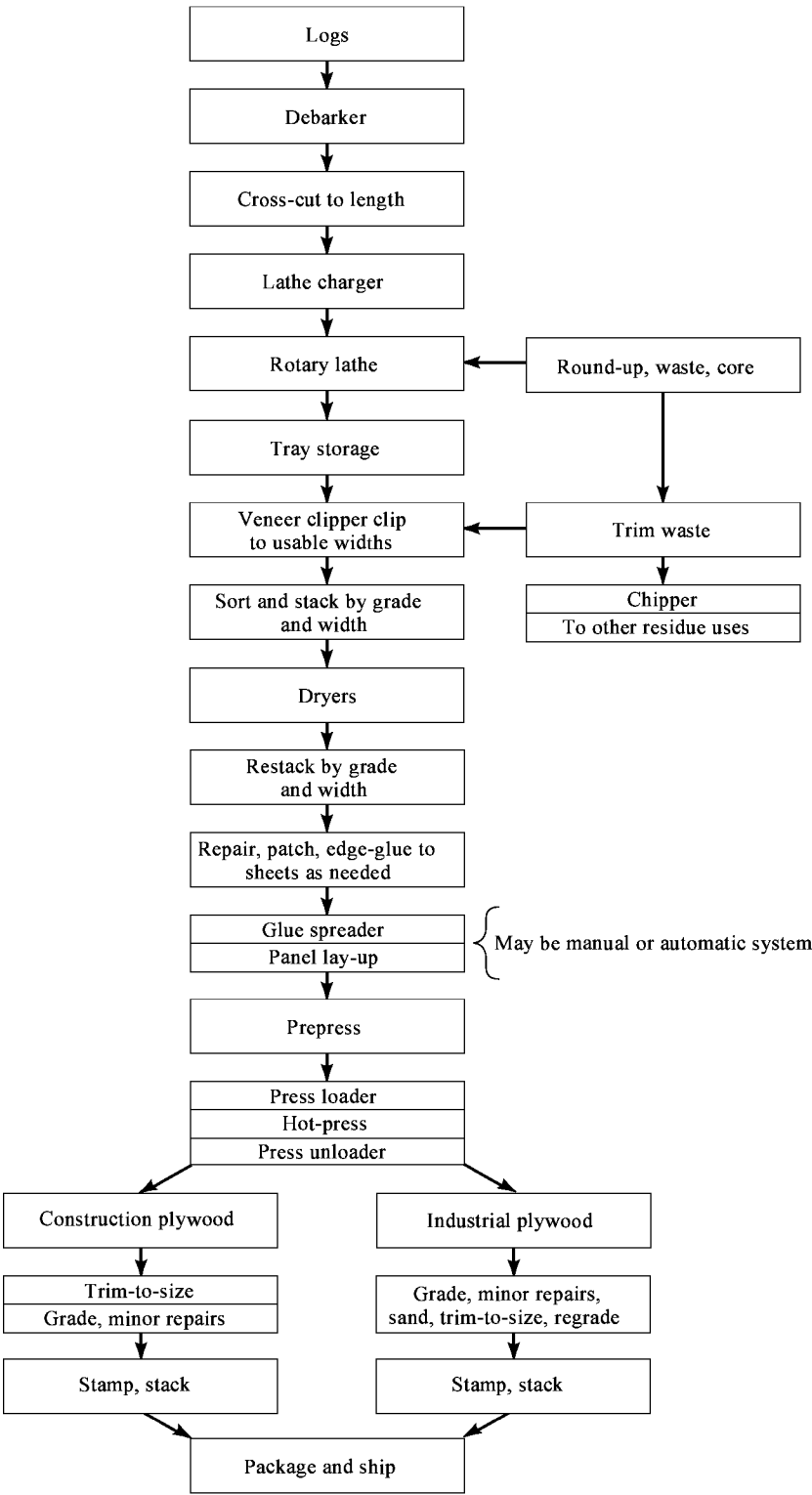


Fig. 3. The softwood plywood process.

After the block is chucked in the lathe, the lathe turns the block against the knife and peels the veneer in a continuous sheet as the knife moves toward the center of the block. When the knife cannot advance further without moving into the metal chucks, the lathe is stopped, the core of the block is dropped, the lathe is recharged, and the cycle repeated.

As the veneer leaves the lathe, it moves into a series of trays, long storage conveyors which provide short-term storage while the veneer moves through the clipper. The clipper is a high speed knife which chops the veneer into sheets which will maximize the yield of usable veneer. Unusable veneer is chopped into chips for use in paper or composites. The veneer sheets move along another long conveyer table where they are pulled and sorted into stacks by grade and width. The stacked veneer is moved to the dryer area where the sheets are placed end-to-end and moved through the long conveyor dryer. Hot air heated by steam, oil, or gas is blown onto the surfaces of the veneer to hasten the drying process. As the veneer emerges from the dryer at a desired moisture content (mc) of 3–5%, a moisture sensor marks those veneers which are not dry enough. These will be redried. The dry veneers are again sorted by grade and width and moved to the splicing and patching area. Here the narrow veneers are spliced together into full-size sheets. Some stacks of narrow veneers are cut in half, from 2550 mm (ca 8 ft.) to 1250 mm (ca 4 ft.) to make veneers for cross-plyies. Some veneers are patched to make higher quality sheets for face veneers. The patching machines punch out defects such as large knots and refill the hole with a similarly sized piece of clear veneer. Then

the stacks of veneer are moved to the gluing and lay-up area. Plywood manufacture is a labor-intensive operation. No other wood-base composite manufacturing operation requires as much hand labor, which is excellent from the viewpoint of providing jobs, but contributes significantly to the final cost of the end-product.

The adhesive used in virtually all softwood plywood has a phenol–formaldehyde (PF) base to provide an exterior-grade, durable, waterproof bond. Thus, most grades of plywood can be used in structural applications. A very small percentage of softwood plywood is made using interior-grade adhesive systems, and this material is used in interior cabinetry, furniture, and shelving.

A typical glue mix for exterior plywood is:

Component	Percentage (by weight)
water	14.65
furafil extender	5.03
wheat flour	3.02
50° caustic soda (NaOH) solution	1.51
diesel oil (defoamer)	0.30
phenol–formaldehyde resin	75.49
	100.00

A typical glue mix for interior plywood is:

Component	Percentage (by weight)
water	69.33
powdered blood–soya mix	15.61
lime (CaO) powder	4.13
50° caustic soda (NaOH) solution	2.39
silicate of soda solution	5.33
phenol–formaldehyde resin	3.21
	100.00

There are two types of gluing and lay-up areas, manual and automated. The manual system is similar to that used in hardwood plywood; surface veneers are laid down and core veneers are passed through a roll spreader where adhesive is applied to both sides. Again, every other core layer is spread with adhesive. Adhesive spreads for western softwoods are usually in the range of 0.88–1.02 kg m² of double glue line (amount of liquid adhesive on both sides of veneer). In the case of Southern pine or hardwood veneers, glue spreads are increased to the range of 1.36–1.52 kg m² of double glue line.

In the automated lay-up system, each layer of veneer (with exception of the top surface veneer) passes under an automatic adhesive application system. This may be a spray application, a curtain coater, or an extruder, each of which is designed to apply a uniform adhesive spread on the upper face of each veneer. After all except the top veneer have been spread with adhesive and laid together, the top veneer is added. The mc of the veneer–adhesive assembly at this point should be about 8°.

The assembled veneers for each panel are placed in stacks equivalent to the number of panels to be pressed in each pressload. Each stack is then placed in a cold single-opening prepress and pressed for a short time to assure transfer of adhesive from the spread faces to the unspread faces. This pressing also provides a small amount of wet adhesion between veneers and facilitates handling of the panels as they are moved into the hot-press. The stacks of panels are moved to a lift on the side of the press. The lift raises the stack and the top panel is pushed into succeeding panel holders on the press loader as the stack passes each holder. The press opens after completion of pressing and curing the adhesive on the current pressload. The press loader then moves the raw panels into the spaces between the hot-press platens. As each raw panel moves into the press, it pushes the newly pressed panel out into an adjacent holder on the press unloader. The press closes and presses the new load for a time sufficient to cure the adhesive. Temperatures are usually about 140°C and presstimes are in the range of 3.75–4.25 min for 12.7 mm (1 2 in.) thick panels. Pressing also results in a slight compression of the veneer, which results in a panel of slightly higher density than the original wood density. Pressed panels move from the unloader into stacks which are set aside to cool for 4–12 h.

There is a hybrid product available which has a veneer back, a layer of PF-coated wood particles, core veneer cross-ply, another layer of wood particles, and a top veneer. This assembly is pressed into a panel, trimmed to size, and sold into the structural-use panel market where it competes with plywood and oriented strand board.

Secondary Treatments and Uses. Depending on grade, panels may receive from no processing to significant processing beyond this point. Structural sheathing panels, a nonappearance grade used as structural wall sheathing, roof sheathing or flooring, may be simply trimmed to size, grade-stamped, and packaged for shipment. Appearance-grade panels will have surface defects shimmed, patched, or filled. Then the panels will be sanded to a uniform thickness, trimmed to final dimension (usually about 1220 × 2440 mm (4 × 8 ft.)), reggraded, and packaged for shipment. Specialty panels may receive additional treatments, such as overlays for concrete form panels, shelving, sign materials, or other uses requiring improved surface properties. Panels for siding might be cut to resemble rough-cut lumber and grooved at intervals to show the appearance of board-on-board siding. Other treatments might include pressure-treatment with preservatives or fire-retardant chemicals for special applications requiring these characteristics.

Economic Aspects. The structural plywood industry now has (ca 1997) about 105 operating mills, representing a significant decrease over the past several years. Production in 1994 was about 17.4 × 10⁶ m³ (2), also representing a marked decrease over previous years. This decrease is a result of several factors, two of the most important being a decrease in availability of suitable veneer logs, especially in the western states, and competition from the newer oriented strand board structural panel industries.

Specifications, Standards, Quality Control, and Health and Safety Factors. APA–The Engineered Wood Association represents the softwood plywood and oriented strandboard industries in the areas of specification, standards, and quality control (QC). An APA product standard, PS1-95 (6), discusses the above areas in detail. The following listing summarizes plywood characteristics covered in PS1-95.

Classification—a review of exposure durability in terms of interior and exterior grades of plywood

Plywood requirements—includes wood species used, synthetic repair requirements, veneer grades, veneer layers and thicknesses, panel grades with respect to end-use, adhesive bond requirements, panel construction and workmanship, scarf and finger-jointed panels, dimensional tolerances, moisture content, and packaging and loading

Testing—includes test specimen preparation, bond durability tests, and structural performance tests. It should be noted that formaldehyde emission tests of phenolic bonded products such as structural plywood are not required because emissions are normally about 0.02–.03 µL L (ppm), well below the previously noted safe level of 0.10 µL L (ppm).

Grademarking and certification—includes certification procedures performed by a qualified inspection and testing agency, and definitions of panel marking requirements

Insulation Board and Structural Fiberboards

Production, Processing and Shipment.

Insulation Board. The panel products known as insulation board were the earliest commodity products made from fibers or particles in the composite panel area. These are fiber-base products with a density less than 500 kg m⁻³. Early U.S. patents were obtained in 1915 and production began soon thereafter. The initial production used wood fiber as a raw material, but later products were made of recycled paper, bagasse (sugar cane residue), and straw. Schematics of the two major processes still in use are shown in Figure 4.

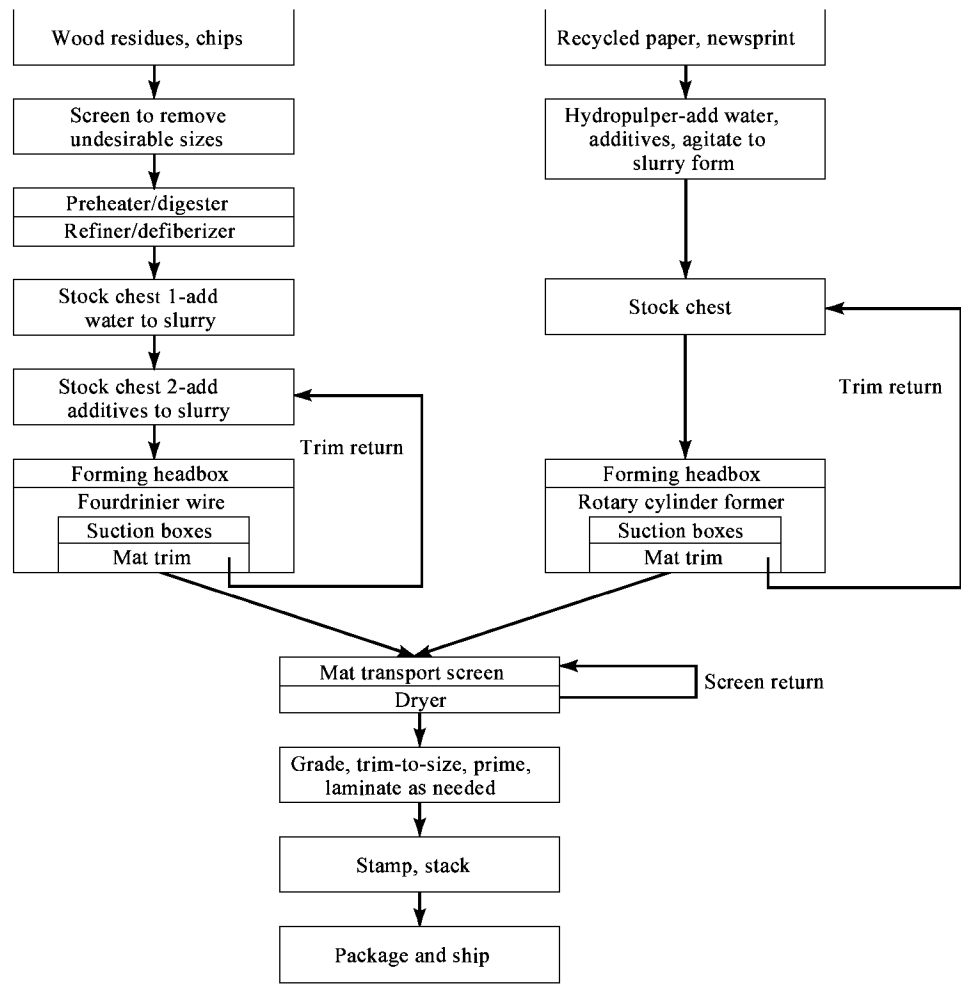


Fig. 4. Insulation board processes.

In the process using wood residues, the raw material (usually wood chips) is passed through a chip washer to remove extraneous materials such as dirt, sand, stones, or metal. The washed chips move to a refiner where the chips pass between rotating steel disks with molded patterns in the faces of the disks. The disk faces are maintained at a small, controlled distance, 0.3–0.5 mm, which results in conversion of the chips into fiber and small fiber bundles as they pass between the disks. The fiber drops into a stock chest, where it is mixed with a large quantity of water. Large paddle mixers are used to keep the fiber in a uniform suspension. This fiber slurry is then moved to a smaller stock chest, where additives are metered into the slurry. Additives may be wax emulsion, starch or PF resin as binder, and alum. Alum is used to change the acidity (pH) of the slurry, which causes the PF resin to precipitate from solution and deposit onto the fiber. One process uses a linseed oil emulsion as a binder. Only small amounts of binder are required in these fiberboards because most bonding is the result of hydrogen bonding, similar to that used in paper-making.

The fiber slurry is pumped to the mat-forming box of a wet-process forming machine. This is a modification of the Fourdrinier machine, widely used in the paper industry. A wide endless wire screen belt, usually about 2540 mm (100 in.) wide, moves under the flow of fiber and water. The water drains through the fiber and through the belt to be recirculated back to the stock chest. The fibers are deposited on the belt at a rate consistent with the speed of the belt to produce a finished panel of the desired thickness and density. The belt then passes over a number of suction boxes, in which a partial vacuum is maintained. The air drawn through the fiber mat draws more water from the mat through the screen and through a series of holes in the top of the suction boxes, further consolidating the mat as well as removing excess moisture. It should be noted that this process uses large volumes of water. Whereas most of the water is recirculated, current regulations concerning clean water have hampered manufacturers, ensuring that in all probability no more of these plants will be built.

The formed mat is carried along on the wire screen and is trimmed to the desired width and length by high pressure water jets. The mats then pass onto another endless wire screen which carries the mats into a long heated drying chamber. The mat may contain 200–300% moisture based on the dry weight of the mat as it enters the dryer. Fans circulate hot air within the dryer, a procedure which removes the majority of the water from the fiber. Also, bonding occurs at points within the mat where individual fibers or fiber bundles touch one another. Some of the bonding is a result of hydrogen bonding at fiber-to-fiber interfaces and some bonding is a result of the presence of additives at these points. As the dry (5–7% mc) mat emerges from the dryer, it is already a useful product, having its own inherent properties and characteristics. These panels are quite low in density, 240–500 kg m⁻³, and a relatively low strength, because primary uses concentrate on insulative abilities and not on high strength. Processes using bagasse or straw are configured similarly to those using wood residues, with minor differences in the fiber preparation steps at the beginning of the process.

The processes using recycled paper fiber are also different at the beginning of the process. Mills using these processes are located near large cities, where large quantities of recyclable paper are available. Of interest in the context of the present emphasis on recycling is the fact that both of the mills producing these products have been operating for well over 50 years.

Paper is placed in a hydropulper, a large water tank equipped with a heavy motor-driven propeller at the bottom. The propeller agitates the water and wet paper and separates the paper fibers into a fiber slurry. Additives are added and mixed as needed at this point. This is a batch process and as such requires more than one hydropulper in order to be processing one tank as another is being emptied of slurry and refilled with water. The slurry is pumped into a box having a large-diameter cylinder covered by a wire screen rotating slowly within it. Fibers are deposited on the screen as water is pumped out of the screen and returned to the process. The fiber mat is carried up and over as the screen rotates and then transferred to a flat wire screen belt which moves the mat across suction boxes and onto the dryers.

Structural Insulation Boards. Structural insulation boards are made by a process similar to that used for insulation board, with the exception of another additive which provides additional weight, strength, and water resistance. The additive is normally asphalt, which is added in a

pulverized form in the additive stock chest. As the asphalt is carried through the dryers on the fibers, the heat of the dryer melts the asphalt, which flows and coats the fibers. When the panels emerge from the dryer and cool, the asphalt hardens, providing added bonding, strength, and water-resistance properties.

Secondary Treatments and Uses. Insulation boards normally have few secondary treatments. Some boards may receive a coating of primer and others may be laminated into panels of several thicknesses. Insulation boards are used for economical, insulative wall paneling, ceiling tiles, bulletin boards, and similar uses. Laminated panels are used for insulative panels, usually as roof decking or insulation under built-up roofing.

Structural insulation boards are used primarily for wall sheathing in constructions where wall diaphragm racking resistance is provided by other means, such as structural panel corner bracing or metal strip bracing. Where allowed by building codes, these methods of construction do provide more economical construction.

Economic Aspects. In 1994 there were 8 operational insulation board producers in the United States. These mills produced about $1.15 \times 10^6 \text{ m}^3$ (2). The number of mills and total production volume have also decreased in this industry, primarily as a result of changes in building codes and availability of other competitive sheathing products. Both wood composite panels and plastic foam sheathings have captured a segment of these markets.

Specifications, Standards, Quality Control, and Health and Safety Factors. Formerly, there was an Insulation Board Institute representing the insulation board industry, but the decline in the market and number of producers has led to its demise. Currently (ca 1997), the industry is represented by the American Hardboard Association (AHA). Specifications and standards are found in American National Standards Institute (ANSI) standard for *Cellulosic Fiberboard* (7). The standard includes descriptions of the various types and classes of fiberboard, as well as requirements for physical and dimensional stability properties. Quality control tests are limited to a few basic strength and stability tests, including bending strength, bond strength, and moisture resistance.

Hardboard and Hardboard Siding

Production, Processing, and Shipment. Hardboards and hardboard siding are fiber-base panel products having densities in the 500–1000-kg m^3 range. Two density classes are made: medium density at 500–880 kg m^3 and high density, >880 kg m^3 . Hardboards are generally thin products, 2.5–9.5 mm in thickness, whereas the siding products are usually 11.1–12.7 mm in thickness.

The first industrial hardboard was developed by W. Mason in the mid-1920s; he found that a mat of wet fiber pressed in a hot press would produce a self-bonded flat panel with good strength, durability, and stability. The product was patented in 1928, trademarked as Masonite, and commercial production began. Over time several other processes for producing hardboards have been developed from modifications of the original process. Brief descriptions of these processes follow and a flow chart of the process is shown in Figure 5.

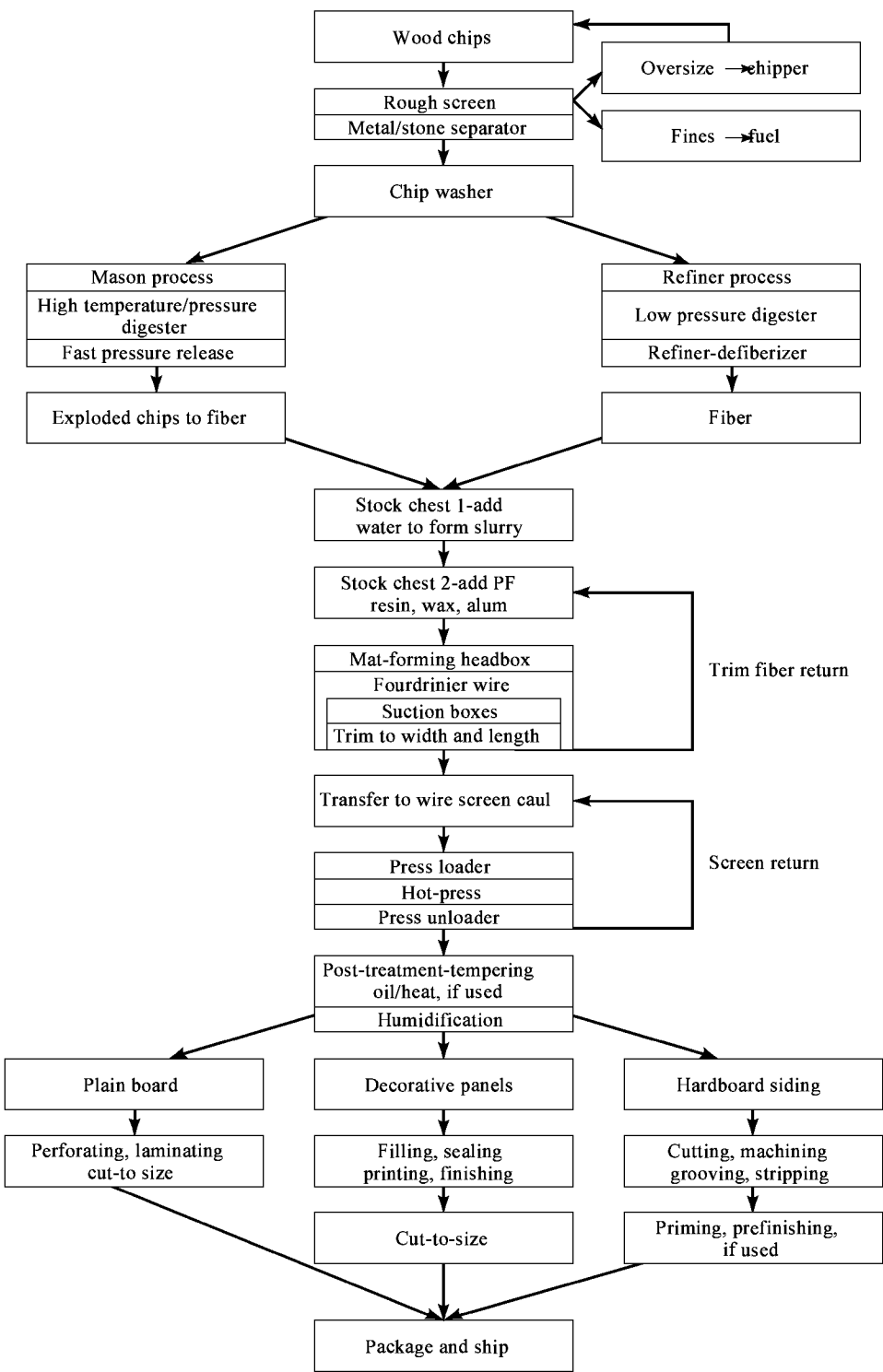


Fig. 5. The wet-process hardboard process.

The Mason (Masonite) Process. In this system, wood chips are sealed in a small steel digester and steamed at high pressure, ca 4.1 MPa (ca 600 psi) for about 1 min. A valve at the bottom of the digester is then opened quickly and the sudden release of the pressure simultaneously defiberizes the chips and blows them out of the digester. The fibers move in a slurry from a stock chest into a Fourdrinier-type mat-forming machine. The speed of the moving wire screen and amount of fiber deposited on the wire screen determine the final density and thickness of the product. The fiber mats pass across suction boxes to remove excess water and are then trimmed to width and length, a size controlled by the size of the hotpress, usually 1270 mm wide and 2490 or 4880 mm long. The mats are transferred from the Fourdrinier wire screen to another wire screen roughly the same size of the mat. This screen carries the mat through the pressing process, where it is removed and returned for another cycle through the process. The mats and screen move into a press loader. When the loader is full, the press opens, discharges the previous load, and the loader moves the next load of mats and screens into the press which maintains a temperature of 190–218°C.

The press closes and the initial pressure and heat force much of the water held in the fiber out of the mat and press as water or steam. The press continues to close until the target thickness and density of the product are reached. The press then maintains this position as the mats continue to absorb heat from the press. The heat converts the remaining water to steam as the temperature in the pressed mat rises above the boiling point of water, 100°C. The steam pressure forces steam through the wire screen under the mat and toward the edges of the press. In this manner the mat continues to dry, and two natural bonding processes take place. The majority of bonding is thought to result from softening and transferring of lignin under moist heat conditions and then drying on adjacent fiber surfaces as the mat dries in the press. The other bonding process is similar to that occurring in paper manufacturer, where adjacent, reactive hydroxyl, –OH, units bond to one another and other reactive sites during the drying process. To aid in removal of water and steam during pressing, it is a common practice to use a breathe cycle. This is done by quickly reducing the press pressure to the point where only a minimal pressure is exerted on the board. Steam escapes with a loud squeal as it passes between the board surface and the press platen surface. After a few seconds, the press is quickly returned to high pressure and remains at proper thickness position for the remainder of the press cycle. It is critical to insert the breathe cycle at the optimum point in the press cycle in order to remove as much moisture as possible in a short time period and then return the press to position before the resin binder cures. Then, when the pressed mats are dry and the press is opened, strong, stable, and durable panels are removed from the press without the requirement of added binder. The original Masonite process, used for many years, followed this procedure. It should be noted that both PF resin and wax were added in small amounts in later years to provide improvements in strength, stability, and durability. The Masonite process is an example of a wet–wet hardboard process; that is, the mat is formed wet and pressed wet.

Wet-Process Hardboard From Refined Fiber. This hardboard is another form of wet-wet hardboard. The principal difference between this process and the Masonite process lies in the method of preparing fiber and the additives used in the boards. Fiber is prepared by processing wood chips through a pressurized refiner. In this machine, chips are metered into a chamber under steam pressure at 520–830 kPa (75–120 psi). As the chips move slowly through the chamber, they are heated and softened by the steam. They then move from the pressurized cylinder (digestor) into the refiner, where they pass between serrated disks rotating at high speed and maintained in close proximity. One or both disks may rotate, depending on the design of the refiner, although most modern refiners use only one drive motor and rotating disk. The rubbing action between the disks separates the chips into fibers, which are actually bundles of fibers mixed with individual fibers. The fibers move out of the refiner into a stock chest, where they are mixed with water into a fiber slurry.

The slurry is pumped into another stock chest, where wax in emulsion form, usually about 0.5–1.0% wax-to-fiber weight, and 1–3% PF resin are added. PF resin is also added on the basis of resin solids-to-dry fiber. Then a small amount of alum is added, which changes the pH (acidity) of the slurry, causing the resin to precipitate from solution and deposit on the fibers. Resin is required in greater quantity than in the Masonite process because only light bonding occurs between fibers prepared in a refiner. The fiber slurry is then pumped to the headbox of a Fourdrinier mat former, and from this point the process is similar to the Masonite process.

Dry-Process Hardboard. Dry-process hardboard is produced by a dry-dry system where dry fiber is formed into mats, which are then pressed in a dry condition. A flow diagram of this process is shown in Figure 6. In this process, wood chips, sawdust, or other residues are refined in pressurized refiners. Wax and PF resin may be added in the refiner or immediately outside of the refiner, in the fiber-ejection tube or "blowline." It is also noted that a small amount of dry-process hardboard is made with UF resin binders. UF resins, because of their inherent faster curing at lower temperatures, can be added only at the blowline or in a blender located after the dryer.

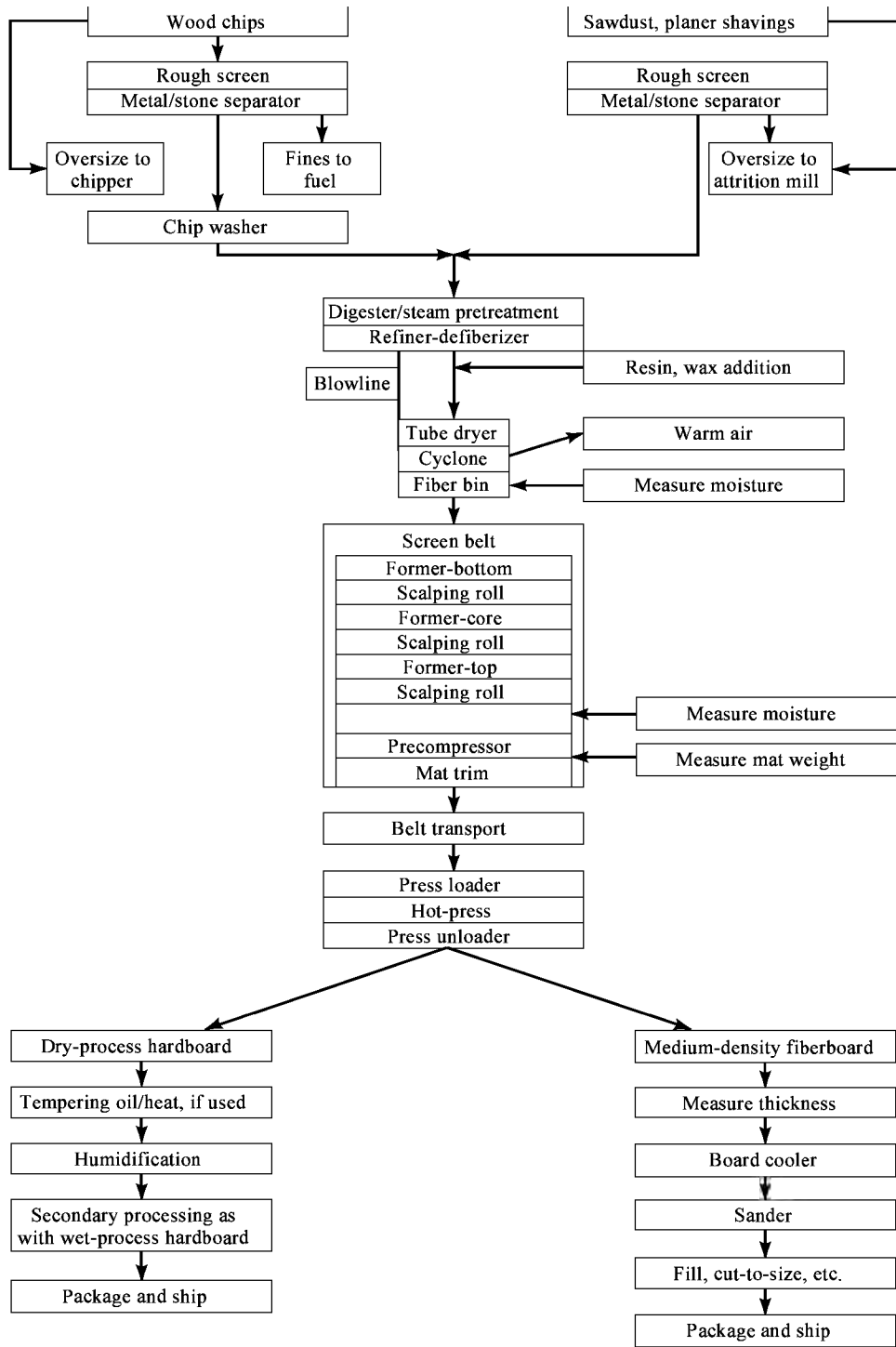


Fig. 6. The dry-process hardboard (and medium-density fiberboard) process.

As the fiber exits the refiner, it moves at high velocity and turbulence into and through the blowline. Blowlines may be 15–30 m (ca 50–100 ft.) in length and 100–150 mm (ca 4–6 in.) in diameter. During passage through the blowline, a sharp decrease in steam pressure occurs. This results in flash-off of moisture and a decrease in temperature of the fiber. Fiber is generally at only 30–40% mc as it exits the blowline. The blowline is an excellent point for addition of wax and resin to the fiber because of the extreme turbulence and mixing which can occur in the blowline. Generally, because little natural bonding occurs in this process, dry-process hardboard requires more wax and resin than wet-process hardboard. Usually 0.5–1.5% wax and 5–7% PF are used (9–11% UF, if used).

From the blowline, the fiber is blown into a tube dryer. Tube dryers are 760–1520 mm (30–60 in.) in diameter and up to 100 m or more in length. Air heated by steam heat exchangers or direct flame is blown through the dryers and moves the fibers in suspension as they dry and pass through the dryer. Temperatures are in the 120–150°C range at the inlet to the dryer and about 70°C at the dryer exit. Fiber moisture should be in the 6–12% range at this point if blowline blending is used. It must be lower if machine blending is to be used later because of water added with the adhesive resin. The resin will not cure in the dryer because of both the short residence time and the cooling effect of evaporating water in the dryer. Where there is moisture in the fiber, this drying and evaporating effect prevents the resin from becoming hot enough to cure.

From the tube dryer, fiber drops into a cyclone which separates the fiber from the moist, warm air. The fiber falls from the bottom of the cyclone into a large bin and in then moved to the mat forming line. If machine-blending of fiber with additives is used, it will occur at this point, immediately before the forming line.

Mat forming is usually accomplished by 2–5 similar machines in sequence, each depositing a portion of the fiber. A plastic or wire screen moves under each forming head, which results in a cloud of fiber falling onto the screen. Under the screen are suction boxes which aid in pulling the loose fiber toward the screen and providing a small amount of mat densification. Immediately past each forming head is a rotating toothed roll called a scalping roll. The scalping roll is adjusted to remove the uneven top of the mat so that the mat will be flat and uniform as it reaches the next forming head. Moisture control is extremely critical, so there will also be moisture-sensing devices situated along the former. These devices provide a measurement of the mat moisture content in order to ensure that moisture is in the proper range at this point and to alert operators of any need for adjustment of moisture. As the mat emerges from the last forming machine and scalper, it should contain enough fiber per unit area to produce the desired thickness and density of the final product. The mat may be 10–20 times thicker than the final board at this point and quite loose in composition. In order to make the mat more compact and to improve the handling capabilities of the mat as it moves toward the press, the mat now moves through a precompressor. This machine is a pair of opposing wide belts which gradually compress the mat to a minimum thickness as it moves between the belts. A thinner mat, 3–5 times the final board thickness, emerges from the precompressor. The mat is then trimmed to standard width and length and from this point is transported on wide belt sections. Usually, there is a device which continually measures the weight of fiber per unit area at this point to ensure that the desired mat weight and uniform cross-panel density are being formed. Mats which are off-weight or uneven in weight will be rejected at this point and the fiber returned to the system for recycling into other mats.

The mats are moved along the line to the press loader. When the loader is filled and the press opens to remove the load of freshly pressed boards, the loader pushes the new boards into the unloader and deposits the load of mats on the press platens. The press closes as quickly as possible to the desired panel thickness. More pressure, as much as 4.8–6.9 MPa (700–1000 psi) is required to press high density dry-process hardboard, because the dry fiber exhibits much more resistance to compression and densification than wet fiber. Press temperatures are also higher, in the range of 220–246°C. No screens are used in the dry-process, but the moisture in the mats requires a breathe cycle during pressing to avoid blowing the boards apart at the end of the cycle. Because no screens are used, the products are called smooth-two-sides (S-2-S), in contrast to the wet-process boards, which have a screen pattern embossed into the back side and are known as smooth-one-side (S-1-S).

Wet–Dry Hardboard. Wet–dry hardboards are a special class of boards in which the fiber is processed through the mat-forming stage in the same manner as in the case of wet-process board. However, an emulsion of a drying oil such as linseed oil is used as the binder and is applied in the stock chest preceding the former. Because this oil requires a long drying and cure time, the wet mats are passed into and through a conveyor dryer, as in the insulation board process. The mats are often sprayed on one surface with the emulsion immediately before the dryer. As the mats emerge from the dryer, every other mat is turned over so that the two oil-sprayed faces are adjacent. This assembly of two mats proceeds to the press loader and press, where it is pressed into one board. The drying oil completes the final cure and bonding action in the press. This process requires extremely high press temperatures, usually about 246°C, and great care must be taken to avoid press fires.

Dry–Wet Hardboard. Dry–wet hardboards are the other possible manufacturing alternative. Fiber is processed as in the dry process up through the mat-forming stage. PF binders are used in this system. Then, as the mats are ready to enter the press loader, a large quantity of water is sprayed onto the surface of the mat. This water saturates the top surface of the mat and is sufficient to raise the total moisture content of the mat to 20–40% mc, depending on the thickness and density of the product. Thinner, more dense products require more water. The press operates at about 200°C, and as the press closes on the mats, the water is heated to steam. The steam quickly heats and softens the mat, requiring less pressure to compress the mat and produces a hard, dense top surface much like a wet-process surface. However, because there is much less total water in the mat in this system compared to the wet-process system, press cycles are significantly shorter, which produces more product per unit of press surface per unit of time.

Hardboard Siding. Hardboard siding is, as the name implies, a hardboard intended for use as an exterior siding material for buildings. These products have been made using all of the previously outlined hardboard processes, using PF or drying oil binders. Two products, one of which is no longer made, used a powdered thermoplastic binder as the durable adhesive. All of these products were and are made to perform adequately as siding materials for the life of the buildings on which they are installed. A normal life cycle, assuming proper installation and finishing, is at least 25 yr.

Several additional facts are worthy of mention in regard to hardboard siding. The majority of the product is made in the medium-density range, between 640–880 kg m⁻³; thicknesses are normally 11.1–12.7 mm. In addition to the normal smooth panels, there are many embossed surface patterns available which simulate rough sawn lumber siding as well as stucco, stone, brick, and shingle patterns. Embossed panels are made by attaching patterned steel plates to the press platens. A mirror-image of this pattern is then embossed into the top surface of the siding as it is being pressed. Almost all siding is preprimed and possibly prefinished before delivery to market. A variation of the siding product is also available as a roofing material which resembles a shingle or shake roof when installed.

Regardless of which of the various types of production processes hardboard panels follow, processing and shipment are generally quite similar for each kind of hardboard panel. Depending on the projected end-use of the panels, some will be tempered or heat-treated immediately after pressing. These treatments provide an added measure of resistance to water, or retard the rate at which the products will absorb water during exposure. The treatment is especially useful in the production of siding or roofing products. Tempering can be a two-stage operation, where the panels are passed through a bath of hot drying oils and then are placed in large racks and passed through a heated chamber. The chamber is heated to about 135°C, and products require 3–4 h to pass through the chamber. Some products may not receive the tempering oil dip, and will only receive the baking or heat treatment. Almost all hardboard products are humidified after pressing and heat treatment, if they are used. Humidification is necessary because the products emerge from the press and heat-treatment, if used, in an almost oven-dry condition and the moisture content of the panels needs to be increased to a level close to that expected in actual use, or about 6–10% mc. Humidification helps to avoid many swelling problems, ie, warping or buckling, which do occur if the panels are installed dry and then allowed to find their own equilibrium in use. The products are generally placed in large racks and passed through a large chamber maintained at high humidity and high temperature conditions.

After humidification, the products are trimmed to size and stacked. The stacks are then moved to the next processing step and many of the secondary treatments of hardboard will take place at the panel production site. These latter may include the following:

For basic hardboard

- packaging and shipment of raw panels
- touch-sanding to close thickness tolerances
- punching to produce perforated board (pegboard)
- sealer coating
- laminating to produce thick panels
- decorative finishes such as wood grain or other decorative designs similar to those used in the hardwood plywood industry
- cut-to-size for specific customer orders.

For hardboard siding

sanding or planing to thickness
cutting or machining decorative groove patterns
cutting into strips for lap siding
cutting into shingle shake patterns
sealing priming with exterior-grade primer
prefinishing with exterior-grade topcoat

Upon completion of the various treatments the products may receive, the panels or strips are stacked, wrapped, labeled, and shipped.

Secondary Treatments and Uses. Because hardboard products are utilized in a myriad of different ways, the variety of secondary treatments used by customers are practically unlimited. Hardboards are used in furniture, cabinets, paneling, doors, toys, and a host of other uses. Post-treatments may include cutting-to-size, finishing treatments with roll-applied patterns, melamine overlays, printed paper overlays, paints, and even some extremely durable and water-resistant coatings used in tub and shower linings or other uses where water contact is frequent and extreme.

A separate mention is merited for a special molded hardboard product. These are made by a process in which either a fiber mat or hardboard panel is placed between two shaped platens and press-molded to a three-dimensional configuration. The most common resulting shape is a doorskin which resembles a wood panel door. The doorskins are bonded to wood frames to make an excellent, attractive, and relatively inexpensive door. This fiber panel molding process is also used to make a wide variety of molded interior linings used in automobile manufacture.

Economic Aspects. In 1994, there were 16 operating hardboard and hardboard siding mills in the United States. Production was 1.535×10^9 (2) in standard hardboard products. These figures do not include the significant quantities of door skin products made, for which production quantities are not tabulated. Production of hardboards has been relatively stable in recent years, considering them as a group. There have been a few new mill closings and a few mill start-ups. In addition, imports of hardboard have also become more common in recent years.

Specifications, Standards, Quality Control, and Health and Safety Factors. The hardboard industry is represented by the American Hardboard Association (AHA). Specifications and standards are contained in several ANSI standards (8–11). These standards define the various hardboard product categories as well as specific product qualities required for each group.

There are five classes of basic hardboard for which ANSI A135.4 (8) summarizes required performance levels of water resistance in terms of water absorption and thickness swelling, modulus of rupture (breaking strength), and tensile strength, both parallel and perpendicular to the surface. These test procedures are found in ASTM D1037-93 (12). For prefinished hardboard paneling, ANSI A135.5 (9) specifies qualifications which are primarily related to the surface finish. These are abrasion resistance, adhesion of finish, fade resistance, gloss, heat resistance, humidity resistance, scrape adhesion, and stain resistance. The AHA ANSI standard for medium-density fiberboard for interior use, ANSI A208.2 (10), has been superseded by the NPA standard for medium-density fiberboard, also ANSI A208.2 (13). This latter standard represents the small amount of hardboard made with UF resin binders which are specifically limited to interior uses, most of which falls into the medium-density group; this standard is discussed in detail herein. The standard for hardboard siding, ANSI A135.6 (11), defines the principal types of siding and the inspection and normal quality control test procedures related to the product. These tests are weatherability of both unprimed and primed surfaces, water resistance, linear expansion, nail-holding tests, bending strength, hardness, and impact resistance.

In addition to the previously noted safety factors associated with these processes, there are additional needs for dust control and ventilation for dissipation of various vapors from pressing, tempering heat treatment, and machining and finishing operations.

Particleboard

Production, Processing, and Shipment. Particleboards are composites made from particles or small pieces of wood or other lignocellulosic residues, in contrast to the fibers used in the various types of hardboard, and the former are bonded together with an adhesive under heat and pressure. There has been a long debate over the origin of particleboards. At least 5 patents between 1880–1930 describe a form of particleboard, long before an economical adhesive was available. When it was discovered that the polymer of urea–formaldehyde (UF) could be used as an adhesive, it quickly became the binder of choice. Particleboards were first made commercially in Europe in the late 1930s and the industry expanded significantly in Germany during World War II. The industrial technology came to the United States after 1945 and has grown steadily in both size and product quality. The primary benefit of the product is that it provides a necessary and useful outlet for the majority of the previously unused wood residues such as sawdust and planer shavings, as well as some chips from sawlogs and veneer logs. Small amounts of urban wood waste and low value logs are also used.

Particleboards are made in thicknesses of 3.2–44.4 mm (1/8–1 3/4 in.), with the bulk of production going into the 12.7–19.1-mm (1/2–3/4 in.) range. Particleboards are made across a wide range of densities, 415–1000 kg/m³, with a generally inverse relationship between thickness and density prevailing. Figure 7 shows a general flow diagram of the particleboard process.

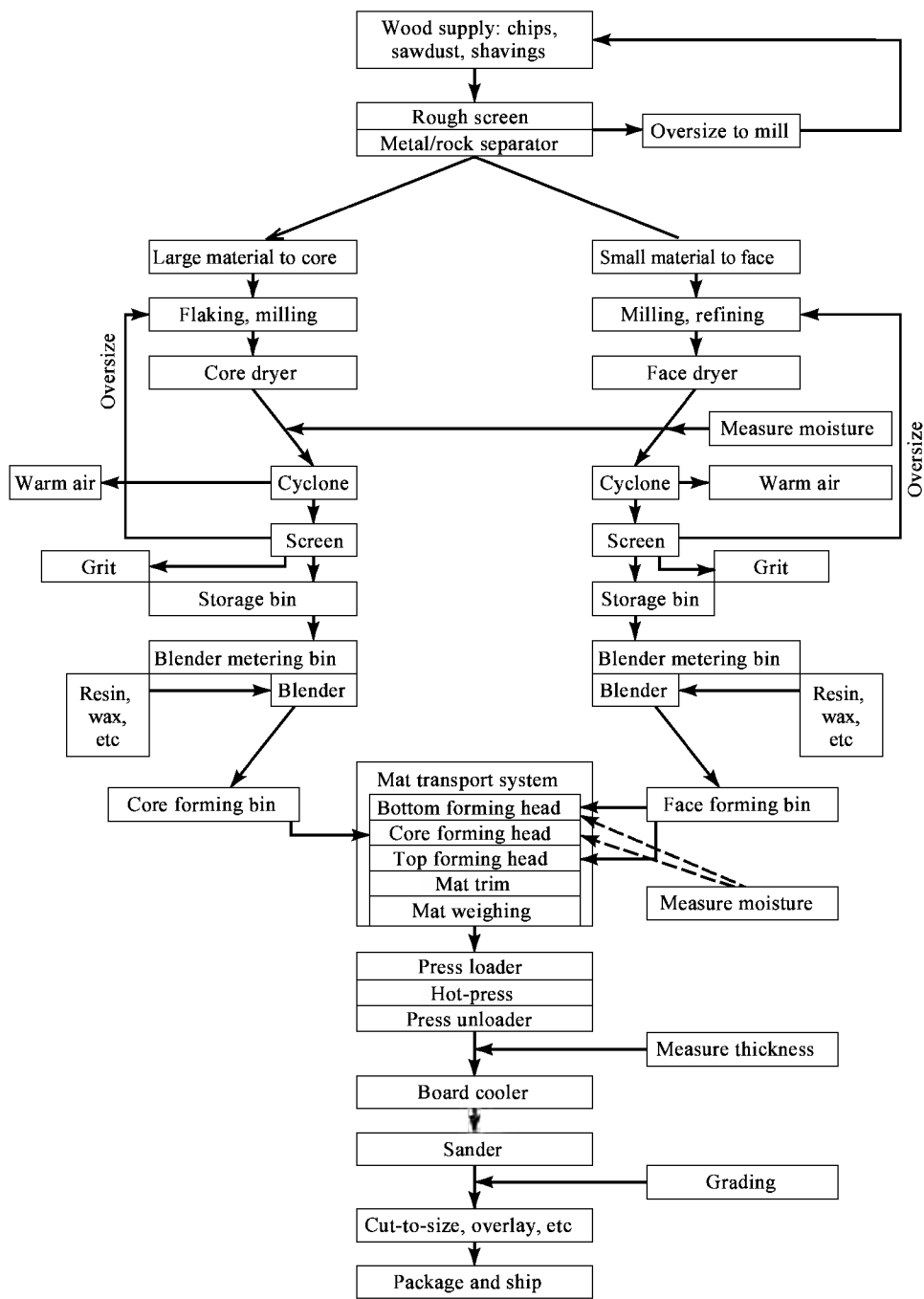


Fig. 7. The particleboard process.

To begin the process, the particulate raw materials are rough-screened to remove oversize materials. The oversize materials may be used as fuel or processed through a grinder and returned to the screen. Also, there should be a unit in this area to remove rocks and trash metal. The material then moves to the milling drying screening area (material preparation area). It should be noted that from this point to the mat-formers, materials usually flow in two streams, one of smaller particles for the surfaces of the board and one of larger particles for the core or middle of the board. Milling generally precedes drying because a better mix of particle sizes is achieved if material is milled when still wet or moist. However, dry milling requires less power and some manufacturers prefer this sequence, even though more fines (small particles) are generated. If dry planer shavings are part of the mix, they will necessarily be milled dry. Another advantage of drying preceding milling is that it allows for separation of grit (dirt, sand, etc) following drying, which reduces wear on grinding equipment and also reduces the chance of fires occurring during milling. Grit removal is not easily done on wet or moist particles. Particulate milling is done by hammermills, impact mills, refiners with special plates, or knife-ring flakers. The knife-ring flaker is a machine designed to make flakes (long, thin, flat particles) from larger residues, such as chips. The other forms of mills use attrition or impact to break up the particles, and a much smaller mix of particles is generally obtained. In a few cases where small logs are used as raw material, these are first debarked and then flaked in large drum or disk flakers. The bark is usually used as fuel.

Drying is almost always done in large, rotating drum dryers. Heat can be provided from direct flame oil, gas, or wood dust burners, or from heat exchangers using these same sources of heat. A modern mill is typically designed to generate as much heat as possible from otherwise unusable wood and bark sources. The wood particles enter the dryers at the hottest (heat inlet) end and proceed through the dryers in a tumbling fashion, carried along by the flow of hot air. Temperatures will be about 540–760°C at the inlet and 200–430°C at the outlet. Dry moisture content of surface material will be at 4–6% mc and core materials at 2–4% mc. After the dryer, particles are separated from the moist, warm air in a series of cyclones. The dryer air then passes through a pollution-control system to remove dust and volatiles, so that the only significant emissions from the dryers are warm air and steam.

A measurement of particle moisture content will normally be taken at the exit of the dryer. This allows the process operators to make such adjustments as may be needed to maintain moisture within the desired range. Various instruments are used, none of which are entirely satisfactory, and periodic hand samples are used in some mills. Considering the importance of moisture sensing and control at the dryers, it is unfortunate that a truly efficient, consistent, and accurate sensing system is not yet available to the industry. The primary reasons for the difficulty of measuring moisture at the dryer exit are the extreme and adverse conditions of heat, dust, and moisture present at this location.

At this point, the dry particles and flakes may pass through a system designed to remove grit. To avoid excessive wear on saws and milling machinery downline and especially for the sake of safeguarding customers' equipment, it is necessary to have a grit (silica) level less than 0.05%, and preferably less than 0.03%. In these days, when more emphasis is being placed on recycling of urban waste woods, grit and metal removal equipment is a necessity for those mills reusing this resource. If the grit is primarily very fine material, it can be removed by a small screen with only small wood losses, and these can be used for fuel. If larger grit is present, more sophisticated equipment is required.

The dry material then passes to the screening area where separations are made, based primarily on suitability as surface (fine fractions) or core (coarse fractions) materials. Oversize materials also is removed and returned to the milling area. After screening, the materials are stored in large bins or silos which provide several hours of running inventory so that the manufacturing process can continue if a breakdown of short duration occurs in the milling drying area. The storage capacity should not be too large, as there is a danger that the raw materials will adsorb moisture and become too moist to allow the process to continue efficiently.

From the storage units, the raw materials are metered at desired and uniform flow rates to blending area. Blenders are normally large tubes employing a rapidly rotating central shaft with mixing paddles attached. Additives such as UF resin, wax, scavenger, and catalyst or buffer are metered into the blender and are mixed by the paddles as the material proceeds through the blender with the desired objective of uniformly mixing the proper amount of additives on all surfaces of the particles. Typical additives used in relative amounts as a percent of dry wood weight are given in Table 2.

Table 2. Additives Used in Production of Particleboard as a Percentage of Dry Wood Weight

Component	Face, wt %	Core, wt %
UF resin ^a	7–11	5–9
wax ^b	0.25–0.75	0.0–0.5
scavenger (urea) ^c	0.4–0.8	0.2–0.5
catalyst (ammonium sulfate) ^d	0.0–0.2	0.0–2.0
buffer (ammonium hydroxide) ^e	0.0–0.3	0.0–0.3

^a Resin solids based on dry weight of wood.

^b Wax solids based on dry weight of wood.

^c Urea solids based on dry weight of wood.

^d Catalyst based on liquid-to-liquid weight with resin.

^e Buffer based on liquid-to-liquid weight with resin.

A small amount of water may also be added to improve blending or to assure proper furnish (blended materials) mc. Actual amounts of resin and wax used in a specific operation are based on amounts needed to produce the desired strength and water-resistance properties. Wax provides, in addition to water resistance, a small amount of lubricity, which aids in moving the materials through the process. The scavenger, normally urea, is added to react with excess formaldehyde from the resin and prevent excessive formaldehyde emissions from the product. In some cases, resin suppliers have been able to formulate resins with extremely low emission potential and thus avoid the additional step of adding an external scavenger. The amounts of catalyst (cure accelerator) and buffer (cure retardant) used must be balanced to meet the requirements of the system. Catalyst is used to hasten the cure of the core adhesive, along with the possible use of small amounts in the surface. Because catalyzed resins are sensitive to temperature, amounts are often changed on a summer-to-winter basis. Careful control is required to balance the need for faster press times and more production, with possible preure of the resin occurring before pressing. Resin durability problems may also be associated with too much catalyst remaining in the board after pressing. Resin suppliers also have proprietary catalysts and buffers which can be added during resin manufacture and in some cases can tailor a resin to the specific needs of the mill so that external catalysts and buffers are unnecessary. Moisture content from the blenders after wax and resin addition should be 9–11% for surfaces and 6–8% for core.

The blended materials, now called furnish, move from the blenders to the formers. A typical mill will have 3–4 formers, one for each surface (top and bottom layers) and one or two for core (middle layer). A mat-carrying plate or caul will pass under the former. This caul will be preweighed and the weight entered on the forming line computer. The caul may be a sheet of steel or aluminum, a wire screen, or a plastic sheet and will generally be the size of the hot-press. On some of the newer systems, the mat-forming and transport system may be a continuous wire screen. The speed of the caul system is such that enough mats may be formed in the time required to press a load of boards, the objective being to maintain the formers in a constant and uniform state of operation. As the caul enters the first former, a layer of fine surface particles is spread uniformly on the caul. The layer will be 20–25% of the panel weight and, depending on the type of former used, may be distributed in such a way that there is a gradation of finer particles on the bottom surface, ranging to coarser particles at the top of this layer. A desirable characteristic of particleboards is a smooth surface, and finer particles on each surface help to achieve this goal. A moisture sensor is often located between the first and second formers.

The caul now moves under the core former(s), where a layer of coarser particles is distributed. This layer represents 50–60% of the board weight. A moisture sensor is often located after the core former(s). A weight sensor may be under the line to sense the additional weight on the caul at this point. Finally, the caul moves under the top surface former and a final layer of fines is distributed atop the core layer. This layer will also be 20–25% of the total mat weight, generally equivalent to the amount formed as the bottom surface. As the caul moves out from the last core former, there is another weight sensor. Given the initial caul weight plus the two weights during forming, the system can easily be adjusted to maintain uniform mat weights by ensuring that the proper amounts are applied by the face and core systems.

After the mat is formed, it may move through a mat precompressor, if this is a feature of the line. Then the mat is trimmed to width and length. If the line does not have the in-line weight sensors noted previously, the trimmed mat will be weighed in total at this point. Mats which are off-weight for the target density and thickness will be dumped and the furnish recycled through the system. Normally, when the line is in uniform operation, very few mats will be off-weight. Mats then proceed to the press loader and into the press. Blended furnish should optimally be formed into mats and be ready to press within 15–20 min after blending. Unless special precautions are taken and special resin formulations, including buffers, are used, material which does not reach the press within 30 min should be dumped and then recycled through the process. It should be apparent how even a short (10–15-min) interruption in operation (downtime) can easily extend to 45–60 min or longer, owing to complications incidental to the interruption.

The pressing operation in a composites facility is a critical one from many standpoints. The amount of high quality board made determines the productivity and profitability of the operation. Thus, it is important to use a press cycle which is as short as possible while producing high quality board and maintaining a consistently uniform operation. But a short press cycle does not ensure high productivity if a number of short interruptions in production occur on each shift. The operators must learn what the optimum speed of operation is in their facility, with the objective of using the press as the slowest point or "bottleneck" in the process. In this way, the press is the only unit in the system which is operating at maximum speed, and thus allows the other equipment to operate at lower and generally more uniform and efficient rates.

Hot presses vary widely in size and number of openings (panels pressed during one cycle). A small size would be 10 openings of 1220 × 2440 mm (4 × 8 ft.). The largest size is 24 openings of 2440 × 7315 mm (8 × 24 ft.). A few single-opening presses may still be found. A recent development is the continuous press, which has two opposing heated steel belts which transport, press, and cure the adhesives as the mat is moved through the press. This is an expensive but highly efficient method of pressing.

The press may operate over a wide range of temperatures, 132–190°C a range of 154–177°C being most common. Higher temperatures allow faster production, but can create other problems if the facility is not specifically designed for these temperatures. Low temperatures are generally only used for thin, high density products which will blister (develop small bubbles) at higher temperatures. A typical press cycle for a 19 mm (3/4 in.) board is as follows, beginning with the press in the open position:

Step	Operation	Time, s
1.	unloading and loading the press	15–30
2.	closing the press to mat contact with platens	5–15
3.	closing the press to final board thickness	30–60
4.	maintaining thickness as adhesive cures	180–210
5.	decompression	10–20
6.	opening the press	10–20

Typically, steps 3–5 require 10–12 s mm of thickness (16–19 s 1 16 in.). It should be apparent that reducing time in steps 1,2, and 6 to minimums is well worth the effort. The combined time in these steps is known as dead time, and mills which have dead times of 60–70 s are at a distinct disadvantage compared to those with dead times of 30–40 s. Another type of press worthy of mention is the steam-through or steam press. In the case of this press, thus far limited to single-opening presses, live steam is forced into the mat through a pattern of numerous small holes in the platen surface as the press is closing. The steam heat almost instantly raises the temperature in the entire board cross-section to a point at which the adhesive cures and thus decreases pressing times dramatically. The process works equally well with particles or fiber, but not as well as with large flakes or strands. One possible disadvantage is that density gradients through the thickness are very uniform and thus the products are not well suited to applications requiring high surface densities. However, for thick, low density products such as door cores, the steam press process has no competitors in terms of overall efficiency.

Desired mat moisture contents of 9–11% for faces and 6–8% for cores are generally accepted. UF resins cure best in a moisture regime of 8–9% mc. An 11% mc surface and 6% mc core are an excellent combination because the high moisture in the surface heats quickly, allowing the surface to compress optimally as well as converting the moisture to steam, which penetrates and heats the core quickly. Too little moisture, as in the case of a dry surface, slows the rate of heat penetration and extends the press cycle. Too much moisture (a more common condition) often requires a significant extension of pressing steps 4 and especially 5 to allow moisture to escape before opening the press. If this is not done, the boards will delaminate (blow) when the press is opened, which may result in the loss of the entire pressload. A severe blow may also cause downtime as the damaged boards are removed from the system. Thus it should be emphasized that success in moisture control is the most important factor in determining the ultimate efficiency and profitability of the operation.

As the panels leave the press area, they may pass through two sensing units. The first is a blow detector, which can locate delaminations that may or may not be seen by visual inspection. Boards with blow areas are marked for removal and then are usually ground up, remilled, and recycled through the process. The second unit is an automatic thickness sensor which, by means of several sensing heads across the board, measures and averages the thickness across and along the board. These thickness measurements and the mat weights taken before the press inform the operators if the product is in the proper thickness and density range and whether or not adjustments need to be made. This is the first product quality monitoring step and it is critical in achieving maximum production of on-grade panel materials.

Panels then move into a cooling device, normally a wheel or rack, where they are held individually and air is circulated between them to remove the majority of heat remaining in the boards after pressing. It is desirable to reduce the average board surface temperature to about 55°C. This temperature is sufficient to complete the cure of adhesive in the core of the board. The heat also helps to redistribute moisture uniformly within the boards, because the board surfaces are drier than the core when the boards come out of the press. Warm boards are normally stacked for several hours to a day to allow for resin cure and moisture equalization.

From these stacks the boards are sanded to final thickness. Most modern sanders can sand to a panel average target thickness ±0.2 mm (±0.008 in.). Sanders are multihead machines, with sanding grits proceeding from coarse to medium to fine. The total thickness removal (sandoff) needed to reach a uniform, smooth sanded surface can be from only 0.75 mm (0.030 in.) on boards from a continuous press to more than 2.5 mm (0.100 in.) on an older multiopening press. Average sandoff on panels from multiopening presses is 1.8–2.1 mm (0.070–0.085 in.). The sanderdust may be recycled into the product, but is more commonly used for dryer fuel. After sanding, boards are visually inspected to remove panels with visible defects such as resin spots, visible blows, or other noticeable marks which affect surface quality.

Boards then proceed to the final processing steps, including trimming to the exact desired sizes, stacking, strapping, and shipping. Some orders may be for cut-to-size of smaller panels, and some mills may also have facilities for filling the surfaces to provide exceptionally smooth, ready-to-finish surfaces.

Secondary Treatments and Uses. Particleboards are the jack-of-all-trades products for interior use panels. They have been and are being used for virtually every conceivable interior use for which panels or strips of material can be used. The principal desirable features of particleboards are flat, smooth, warp-free, stable, and cost-efficient panels. They have good strength, stability, and screw and fastener holding abilities. They are almost ideal substrates for most finishing or overlay systems. They are available in many thicknesses and panel sizes and can be easily worked with good wood-cutting, molding, shaping, drilling, and vee-grooving equipment. To put their versatility into the perspective that is their due, it would be safe to say that the vast majority of homes and businesses in the United States contain a number of products in which the basic substrate is particleboard.

The majority of particleboard is used in furniture and cabinetry. A significant amount is used as floor decking in manufactured homes and as underlayment in conventional homes. Particleboard is used as shelving in homes and industrial businesses, as door cores in solid core doors, in stair stepping, door frames, and a host of uses requiring small flat parts as a starting point.

A small amount of particleboard is made with a fire-retardant treatment for use in locations where codes require this material, as in some offices and elevators. Particleboards receive overlay and finishing treatments with ease. Wood veneers, melamine overlays, printed paper overlays, vinyl overlays, foils, and direct grain printing can all be done quite simply. A small amount of particleboard is also made in the form of shaped, molded articles such as furniture parts, paper roll plugs, brush bases, and even toilet seats. There is another small increment of particleboard made by the extrusion process. These products are made in small captive operations owned by furniture manufacturers which consume all of this production in their furniture. The extrusion process differs from conventional flat-pressed particleboard in that the wood furnish is forced between two stationary heated surfaces. The mats are formed from one edge and this edge is alternately formed and pushed between the heated platens, which are maintained at a distance equal to the thickness of board produced. This is an old, slow, small-scale process, but is still in use in at least one location.

Economic Aspects. In 1994 there were 45 producers of conventional particleboard in the United States. These mills produced 8.205×10^6 m³ of product (2). The real growth of the industry occurred in the period from the 1950s through the 1970s. The industry continues to grow slowly, with one new mill on the average being added annually. The major growth worldwide in traditional particleboard markets is occurring in the medium-density fiberboard area.

Specifications, Standards, Quality Control, and Health and Safety Issues. The particleboard industry is represented by the National Particleboard Association. Specifications, standards, and quality control (QC) procedures are outlined in *Particleboard*, ANSI A208.1-1993 (14). There are five grades of particleboard, with 1–4 subgrades within each group differentiated by physical property differences. These are requirements for bending and stiffness strength, bond strength perpendicular to the surface, hardness, screwholding properties, linear expansion, and formaldehyde emission. All of these properties are included in a normal QC regimen of tests. Physical properties test procedures are outlined in ASTM D1037-93 (15). Formaldehyde test procedures are found in ASTM E1333-90 (4).

In addition to the always-present concerns for worker safety around hot or moving machinery, additional measures are warranted by the dusty conditions that generally prevail around a particleboard mill. Great care is exercised to address the conditions through the use of cyclones and baghouses to trap dust and prevent its release into and around the mill environment. However, constant clean-up is necessary to prevent excessive dust buildup, which left unchecked can become a fire and explosion hazard. Most equipment which could start a fire or cause an explosion if foreign material were to create a spark within the system is protected by spark-detection and automatic fire suppression equipment. Conveyors are equipped with pressure-release hatches to reduce the danger of fire and explosion traveling throughout the system. Some workers in dusty areas wear dust masks. Most workers wear safety

glasses and often make use of hearing protection devices, safety shoes, and hardhats. Advances in UF resin technology and use of scavengers have made concerns about formaldehyde exposure almost nonexistent in both mill and product use environments.

Medium-Density Fiberboards

Production, Processing, and Shipment. Medium-density fiberboards (MDF) are panels made of fibrous raw material and used in most of the same applications as particleboard. MDF products generally have more smooth surfaces and edges than particleboards and are thus preferred for some uses, even though the manufacture of MDF is more costly and the product is significantly more expensive.

Historically, MDF was developed in the United States in the 1950s. After a slow start, the industry grew most rapidly in the 1970–1990 period and has grown slowly in the United States since then, giving indications of significant potential for expansion in the 1998–2000 period. In recent years, other countries throughout the world have also recognized the favorable traits of MDF and many more plants have been and are now being built, particularly in Europe, Asia, and the South Pacific areas.

The manufacture of MDF, with a few exceptions, duplicates the manufacture methods for dry-process hardboard, described at length herein. One exception to it is that most MDF is made in the medium-density range, 640–800 kg m³ although small amounts are made at lower or higher densities. Second, the vast majority of MDF is made with UF resin adhesives with resin requirements in the 7–11% range, and wax is usually added at the 0.50–0.75% level. A small amount of exterior-grade MDF is made with isocyanate resin.

The major difference from dry-process hardboard manufacture is in the pressing area. Because MDF is usually a lower density product, lower pressing pressures are needed, and presses are not quite so robust compared to the dry-process hardboard presses. Also, some early MDF presses used primary steam heating and secondary heating systems based on radio-frequency (RF) heating. The use of RF heating allowed the center of the board to heat almost as quickly as the surface, and this occurred during closure of the press. Thus, the heated core, would compress almost as easily as the faces, which resulted in a quite uniform density profile through the thickness of the board. This then produced the appearance of a uniform, smooth edge when cut or machined.

The other major benefit of RF heating was in reduced presstimes. A typical steam-heated MDF press was operated at about 163°C. Presstimes, not including deadtime, for 19-mm (3/4 in.) board would be about 7 min. With RF, this time could be reduced to about 5 min. It will be noted that these presstimes, even with the use of RF, are longer than those required for particleboards and this, in addition to the more costly base fiber and the higher resin requirements, explains much of the manufacturing cost differential between MDF and particleboard.

After pressing, the MDF process basically duplicates the particleboard process with the steps of cooling, sanding, trimming, cut-to-size, stacking, strapping, and shipping.

Secondary Treatments and Uses. MDF competes with particleboard in virtually every application, especially in furniture and cabinetry. However, MDF is not used as floor underlayment or decking except on special request. The popularity of MDF is the result of the smooth surfaces and smooth-cut or machined edges. For applications such as direct grain printing, laminating with thin overlay papers or foils, application of smooth, gloss finishes, or uses requiring smooth, highly machined edges, MDF is the product of choice.

The exterior form of MDF is used in special applications requiring durability and resistance to water or weather exposure. Highway signs would be an example of this use of exterior MDF. It is an extremely expensive product and thus is used only for special applications requiring its special properties. Another example of use would be where a customer would be willing to pay the additional cost to use a composite which has the exceptional qualities of MDF, but also has virtually no formaldehyde emissions.

Economic Aspects. In 1994, there were 14 producing MDF mills in the United States. These mills produced 2,240 million m³ of product (2). The market for MDF in the United States is fairly well saturated at this time and for this reason the industry is expanding only slowly. However, as noted herein, the world market is still growing rapidly and the manufacturers are building to satisfy this market. Currently, some U.S. manufacturers are also exporting MDF into these markets.

Specifications, Standards, Quality Control, and Health and Safety Aspects. The MDF industry is represented in the areas of specifications and standards by the NPA. Specifications and standards are found in *Medium-Density Fiberboard (MDF)*, ANSI A208.2-1994 (13). There are four classes of MDF, three classes divided by density and one for exterior-bonded MDF. Primary properties outlined in the standard and used in QC tests are bending strength and stiffness, bond strength perpendicular to the surface, screwholding properties, and formaldehyde emissions. Strength tests are described in ASTM D1037-93 (12) and formaldehyde emissions in ASTM E1333-90 (4).

The health and safety issues outlined herein for particleboard also apply to MDF. A special note should be made of the fact that, because the MDF raw material is of dry fiber base, there exists in MDF a large component of very small, broken, dust-like wood fibers. These contribute to the dust concerns in the manufacturing areas, requiring excellent dust-control systems, good housekeeping, and personal protection.

Waferboards and Oriented Strand Boards

Production, Processing, and Shipment. The waferboard and oriented strand board (OSB) industries are based on the use of special forms of wood flakes generated from small logs. A flake is a long, flat section of wood which may be 25–100 mm (1–4 in.) in length and in the grain (longitudinal) directions of the wood. Thickness may be 0.25–1.00 mm (0.010–0.040 in.), and width is usually variable. Normally, a good flake has a length-to-thickness ratio of at least 100.

Wafers as used in the industry were large, flat flakes of about 0.6–1.0 mm (0.025–0.040 in.) in thickness, 38–50 mm (1 1/2–2 in.) in length, and 13–50 mm (1/2–2 in.) in width. Strands are long, narrow flakes of about 0.6–1.0 mm (0.025–0.040 in.) in thickness, 75–100 mm (3–4 in.) in length, and 6–25 mm (1/4–1 in.) in width. Variation will occur around these dimensions, but these will apply in the case of the majority of the waferboard and OSB products.

Flakes and flakeboards have long been a source of attention and interest to the industry, because excellent strength and stability can be achieved with boards made from flakes. However, the requirement of solid wood as a starting material and uneven swelling properties of the boards prevented them from becoming an economical or desirable product for use in many particleboard applications. Nevertheless, at least seven 1950s era particleboard plants operated on large flakes as part or all of their raw material. Decorative paneling was probably the most successful product of this production, but eventually all went out of production because they could not compete in the standard particleboard markets. In 1955, a plant was built in Canada utilizing the technology of an earlier U.S. mill: large flakes or wafers bonded with powdered PF resin adhesive. The market was low cost, exterior sheathing panels called Waferboard for barns, sheds, fences, etc. Time spent in exposure proved that the panels were entirely satisfactory for these uses; in fact, many of these structures are still in use today. Over the succeeding years, a few more plants were built, mostly in Canada.

Then, in the early 1980s the concept of OSB was realized in the construction and operation of large-size mills. OSB is a panel product made from wood strands and somewhat like plywood in that the strands on the two faces are oriented in the long direction of the panel and the core strands are oriented in the cross-panel direction. The use of orientation yields panels having excellent directional properties, much like plywood, and thus an excellent and economical structural sheathing material is created.

The manufacture of waferboard and OSB has many of the same process steps as particleboard, but adapted to the special needs of producing an exterior quality panel with large wafers or strands. This discussion focuses on OSB, because waferboard has been almost entirely replaced by OSB and most of the early waferboard mills have now been converted to production of OSB. The OSB process is outlined in Figure 8.

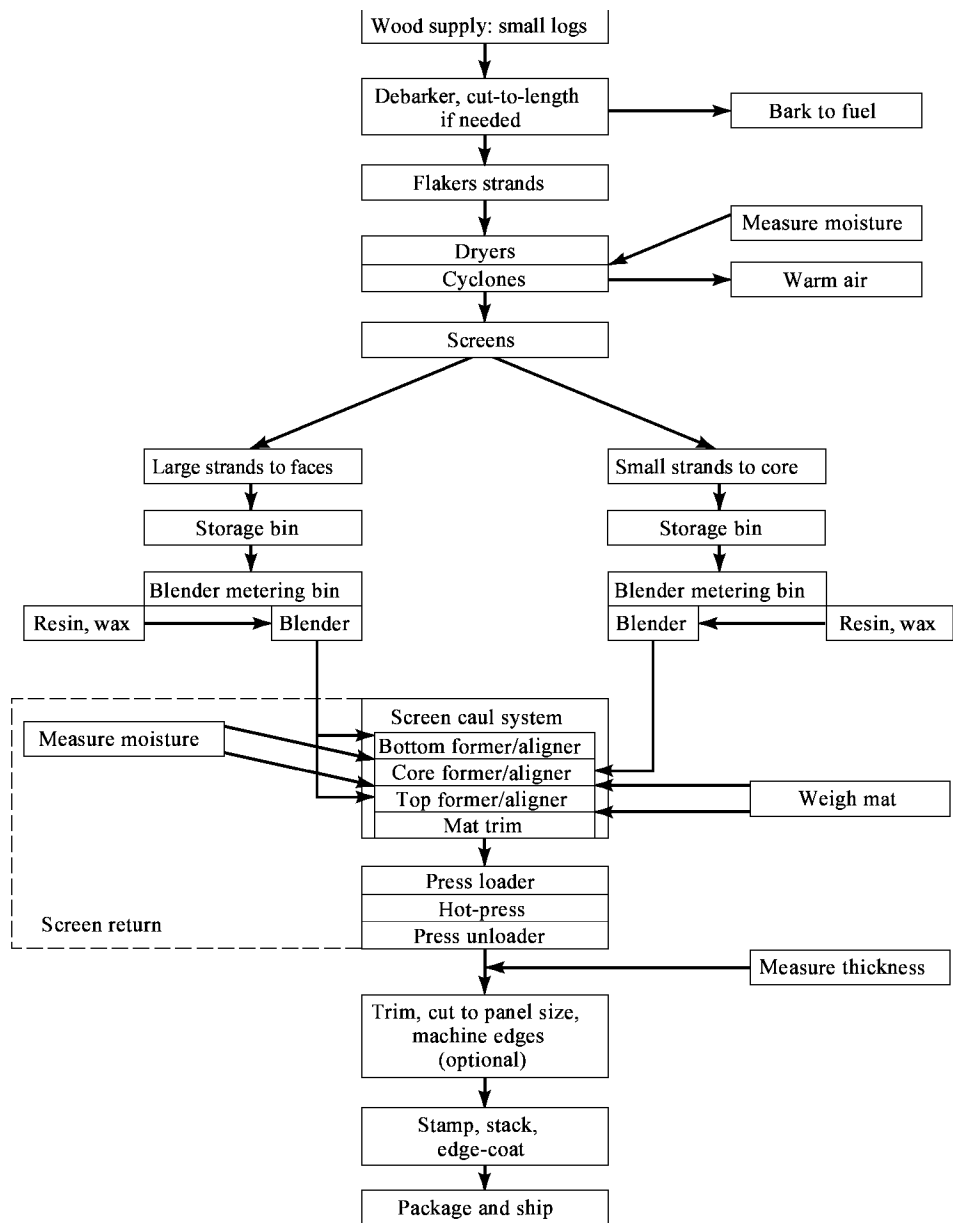


Fig. 8. The oriented strand board process.

The process begins with small logs of almost any species. Aspen was the preferred species for years but now the industry has spread well beyond the range of aspen growth and any species which can be converted to strands has been used. High density hardwoods such as oak, hickory, and hard maple are avoided where possible, because of the difficulties in flaking. The logs to be processed are debarked and the bark used for fuel. The debarked logs are placed in soaking tanks which will be heated in winter to thaw and heat the frozen logs. As the logs reach the end of the soaking tanks, they are removed and prepared for the flakers. Flakers are of two types, the first using huge rotating disks containing a number of long knives extending outward from the axis of the disk. These knives protrude from the face of the disk by the desired flake thickness. There are also small, sharp spur knives protruding from the disk at intervals equal to the desired flake length. Log sections are fed into the disk and each knife removes a slice from the log. The spur knives cut the slice into many separate lengths, and these pass through an opening in the disk. As the slices pass through the disk, they encounter special metal shapes which bend, buckle, and break the slices into narrow strands which fall into a conveyor below and behind the disk. Another type of flaker is a revolving ring flaker. The ring is about 600 mm (24 in.) in depth, with many flaker knives and spur knives protruding inward on the inner periphery of the ring. Logs or log sections are pressed against the inner periphery as the ring rotates. Slices of wood are removed and are again broken into strands as they pass through the ring and fall into a transport conveyor. Recent developments have made it possible to flake whole logs one section at a time in both of these flaker types.

The strands move through large drum dryers which reduce the moisture content to about 2 or 6%, the difference being whether liquid or dry resin is to be used. Because a desired moisture level into the press is about 6–7% mc, a liquid resin adds water to the system and requires a lower flake moisture than a dry resin.

From the dryer, the strands are screened to remove fines and small particles, which would detract from board quality and economy. These fines are burned for fuel or possibly sold to a nearby particleboard mill as raw material. The larger strands are used as surface material and the smaller strands are core material.

The strands are metered from the bins into the blenders, which are large rotating drums which tumble the flakes as they move from the feed to the exit end. Resin adhesive and wax are applied; if dry, powdered PF is applied, this is accomplished by a simple metering system which meters the powder at a rate proportional to strand flow rate and at a level of 2–2.5% by weight of dry wood. Other liquid additives are normally applied by centrifugal disk atomizers rotating at high speed and generating a mist of additives which attach to the surfaces of the tumbling strands. Wax is applied at a 1–1.5% rate to provide water repellancy in the finished product. Liquid PF adhesives are applied to rates of about 3.5–4.0%. Another binder being used in some locations is isocyanate resin. It may be used in the entire board or it may only be used in the core of the board with PF-bonded surfaces. Usage rates are 1.5–2.0%. Isocyanates have several desirable features, such as excellent bonding abilities, high moisture tolerance, faster curing than PF, and lower adhesive usage than PF. Less desirable attributes are press sticking potential, high cost, and workplace health and safety concerns. Press sticking can be avoided by use of isocyanates in the core only where faster cure rate also helps to offset the added cost.

Strands move from the blenders to the formers. Mats are formed on wire screen cauls. The major benefits of screen cauls are long working life and as an aid to removal of moisture in steam formed during and immediately after the press cycle. This dramatically reduces the tendency of these large strand products to blow after pressing. An unintentional benefit was the embossed screen imprint in the board, which when placed screen-side-up on roof sheathing installations, provides a slip-resistant walking surface for workers. As the screen-cauls pass under the formers, the lower surface strands are

placed on the screen-caul such that they are generally aligned with the long direction of the screen. This is done by dropping the strands between rows of closely spaced, rotating metal disks located just above the screen caul. The core strands are then laid down so that their lengthwise dimension is generally perpendicular to the surface strands. Finally, the top layer of strands is applied to the mat, also in the long direction of the screen and mat.

Mats then move to the press loader and into the press. Presses in OSB plants are usually very large, with panels of 2440 × 7320 mm (8 × 24 ft) a common size. Newer plants produce panels of 3660 × 7320 mm (12 × 24 ft). Presses will produce 8–16 panels of these sizes in one pressload. Press temperature are 204–218°C for pressing PF-bonded panels, but temperatures as low as 190°C can be used when isocyanate adhesives are used. Press times, not including deadtimes, with PF-bonded panels are about 14–15 s/mm of thickness. With isocyanate adhesives, presstimes are reduced to about 12 s/mm of thickness. The presses are equipped with elaborate position-control systems so that all panels will be close to desired thickness as they exit the press. Panel densities will range from 575–675 kg/m³, depending on wood species used and end-product use requirements. The objectives are to reduce panels to minimum densities consistent with good quality and end-use requirements. This extends the resource, as well as easing handling in the field.

After the press, the panels are immediately cut into ordered sizes, the most common of which is 1200 × 2440 mm (4 × 8 ft.). Panels are then printed with appropriate use information, stacked, and usually edged coated with a moisture-proof barrier coating prior to banding and shipping. The edge coating provides protection from moisture damage on panel edges during shipping and field installation of the panels.

Secondary Treatment and Uses. The vast majority of OSB panels are used "as is," without further processing or treatment. Primary uses are as wall and roof sheathing, floor decking, and other construction panel uses in home and commercial construction. OSB products are effectively filling in for the decline in plywood production. Small amounts of OSB are used in furniture, primarily as frame stock, and in other uses in which plywood might be used.

The major secondary use of OSB has been the production and marketing of an OSB exterior wall-paneling or siding. This was done by applying a PF-resin impregnated paper on the top of the strand mat immediately before pressing. The paper overlay was then bonded to the mat during pressing and embossed with a rough-sawn wood grain pattern in a manner similar to that used for hardboard siding. Panels were then sawn into strips, coated with a paint primer, and sold as lap siding. Another competitive product was made with a layer of adhesive-coated fiber bonded to the top of the strand mat in the press. This product was also embossed in the press and then primed and sold as lap siding strips.

Economic Aspects. There were 31 OSB plants in the United States in 1994 and these produced 6.625 × 10⁶ m³ of OSB products (2). This industry is growing rapidly in both Canada and the United States. In fact, many of the composite mills currently under construction are designed to produce OSB or similar products based on strands. Outside of North America, where building practices are not yet extensively utilizing the distinct advantages of the stud wall and plywood OSB sheathing, there are only a few operating OSB plants. There are also small export markets for OSB products in Europe and the Far East.

Specifications, Standards, Quality Control and Health and Safety Aspects. The OSB industry is represented by the APA—The Engineered Wood Association in areas of specifications and standards. These are outlined in APA *Performance Standard for Wood-Based Structural-Use Panels*, PS2-92 (15). The standard defines specifications, qualifications, and tests for panels of any type which are to be used in structural-use applications. Thus, structural plywood, OSB, and all other similar products are considered to be covered by this standard. The major areas discussed in the standard are product classification, general requirements, and product evaluation and qualification. Twenty test methods are outlined and can be placed in the following general groups: bending and load testing, fastener holding, moisture and dimensional stability, mold and bacteria tests, probe tests, and cyclic durability tests.

Health and safety factors in the OSB industry are similar to those in other composite mills. Worker safety cannot be stressed too highly, a main component of which is to develop an awareness in the workers to be prepared for danger at all times. Many accidents are the result of a moment of careless or unthinking activity on the part of the person injured or another nearby person. An area of special concern in some OSB mills, those using isocyanate adhesives, is awareness of the toxic nature of this adhesive in the uncured state and the requirements of personal care, housekeeping, and ventilation and air handling in the process areas from blending through pressing.

Structural Composite Lumber

Production, Processing, and Shipment. Structural composite lumber is a group of composite or veneer products which can be used in place of structural lumber in many applications. These products and markets have been developed to fit specific needs of the construction industry where solid lumber of desired grades is no longer available or at least, very difficult to obtain. An important factor in the acceptance of these solid lumber substitutes is their uniformity. While composites in general may not be as strong as clear, straight-grain solid wood, they may be significantly stronger than knotty or angled-grain lower grades. Also, the variation in properties is much less than in run-of-the-mill solid lumber. Thus, in the design of wood structures for which a composite with known and predictable properties can be selected, as compared to a batch of lumber, in which the worst piece defines the batch, there is no question which material should be selected.

There are three distinct types of structural composite lumber available. The manufacture of each is discussed herein with emphasis on interrelationships between the products.

Laminated Veneer Lumber. Research and development on laminated veneer products resulted in manufacture of a marketable lumber substitute in the late 1960s. The product was made of many layers of 2.5–3.2 mm (1/10–1/8 in.) veneer bonded with PRF adhesive in a long continuous press. Panels 1220 mm (4 ft.) wide and 38–50 mm (1.5–2.0 in.) thick of endless lengths were made. Ends of the 2440 mm (8 ft.) long veneers were staggered through the thickness with a short overlap at the ends of veneer to provide a lap joint to transfer loads. Since that time, other manufacturers have entered the laminated veneer business. Some thicker veneers with various endjoints including butt joints and fingerjoints have been seen. Properly made, these materials approach the load carrying abilities of high grade solid lumber. Because of their reduced variability in strength, these laminated veneer products compare very favorably with the highest quality lumber.

Veneer-Composite I-Beams. One of the earliest applications of laminated veneer lumber was as flange stock in a veneer-base I-beam. Narrow strips, 50–100 mm (2–4 in.) were cut from the 1220 mm (4 ft.) wide panel and a groove about 9.5 mm (3/8 in.) wide and deep was cut in the surface of two of these strips, and the two strips were placed in a machine with a set distance separation between them. As the strips moved through the machine, a PRF adhesive was metered into the grooves. Then cross-cut sections of 9.5 mm (3/8 in.) plywood were placed in the grooves end-to-end to make web-stock for the I-beam. The strips were pressed together to seat the plywood in the bottoms of the grooves, and the assembly came out of the machine as a completed I-beam. As an example, a nominal 2 × 12 in. lumber replacement, actually 38 × 286 mm (1.5 × 11.25 in.), might have flanges 63 mm (2.5 in.) wide × 38 mm (1.5 in.) thick and with a web of 229 mm (9 in.) to duplicate the 286 mm (11.25 in.) dimension of the solid lumber. In this way, beams of almost any grade and size can be made by changing the width of the flange and with width of the web material. In recent years, OSB has replaced much of the plywood as web material in these I-beams.

Laminated Strand Products. The most recent developments in the family of wood-based composites are a group of laminated strand products, made with strands oriented in the long direction of the product and marketed as structural composite lumber. One product is made with long, narrow strips of softwood veneer. The strips or strands are about 2.5 × 13 × 600 mm (0.1 × 0.5 × 24 in.), coated with a PRF adhesive, and pressed under heat and pressure into large blocks. After the resin is cured the blocks are resawn and planed into lumber dimension stock.

Another product is made with strands produced by flakers, having dimensions of about 0.8 × 13 × 300 mm (0.030 × 0.5 × 12 in.). The strands are dried and then coated with isocyanate adhesives. The strands are formed into mats with unidirectional orientation and pressed in a steam-through press. The large panels are then sawn into dimension lumber sizes.

Secondary Treatments and Uses. The structural composites lumber group requires little secondary processing to be used in the various construction areas for which they are intended. The products are being used in most of the applications formerly and almost solely filled by large structural timbers and large-dimension lumber. Some of these uses are as floor joists and roof trusses, both of which are ideal applications for the composite I-beam; building foundation timbers; garage door headers, as well as headers for large doors and windows; scaffolding planks and similar uses.

Economic Aspects. In 1994, there were 10 manufacturers of various types of structural composite lumber with a production volume of about 0.820 × 10⁶ m³ (2). Several more plants are now in production or under construction. Given the reduced availability of solid lumber in large structural sizes, this is another area of composite products having a promising future.

Specifications, Standards, Quality Control, and Health and Safety Aspects. In general, structural composites lumber products are tested and rated in the same manner as structural lumber. Strength in bending stiffness is possibly the most important characteristic, because this property defines the basic load-carrying capacity of the product. Tensile strength is also important, because this defines load-carrying under a tension stress, such as in the bottom chord or flange of an I-beam. Terminology, procedures for determining allowable design stresses, and quality assurance requirements are set forth in the ASTM *Standard Specification for Evaluation of Structural Lumber Composite Products* (D5456-93) (16).

Health and safety concerns in the structural composite lumber industry are similar to those in the other composite industries. Special care is required in worker awareness, worker protection equipment, dust and vapor control, and general housekeeping.

BIBLIOGRAPHY

"Plywood" in *ECT* 1st ed., Vol. 10, pp. 860–870, by J. G. Meiler, Coos Bay Lumber Company; in *ECT* 2nd ed., Vol. 15, pp. 896–907, by C. B. Hemming, U.S. Plywood-Champion Papers, Inc.; "Laminated Wood-Based Composites" in *ECT* 3rd ed., Vol. 14, pp. 1–41, by J. A. Youngquist, U.S. Department of Agriculture.

1. *Manufactured Home Construction and Safety Standard*, Regulation 24CFR 3280.208, U.S. Department of Housing and Urban Development (HUD), Washington, D.C., Feb. 11, 1985.
2. T. Sellers, Jr., *Panel/World* 37(2), 12–13 (1996).
3. *Hardwood and Decorative Plywood*, ANSI HPVA HP-1-1994, American National Standards Institute, New York, 1994.
4. *Standard Test Method for Determining Formaldehyde Levels from Wood Products Under Defined Test Conditions Using a Large Chamber*, ASTM E1333-90, American Society for Testing and Materials, Philadelphia, Pa., 1990.
5. *Standard Test Method for Surface Burning Characteristics of Building Materials*, ASTM E84, American Society for Testing and Materials, Philadelphia, Pa., 1975.
6. *Construction and Industrial Plywood*, APA Product Standard PS1-95, APA—The Engineered Wood Association, Tacoma, Wash., 1995.
7. *Cellulosic Fiberboard*, ANSI AHA A194.1, American National Standards Institute, New York, 1985.
8. *Basic Hardboard*, ANSI AHA A135.4-1982, American National Standards Institute, New York, 1982.
9. *Prefinished Hardboard Paneling*, ANSI AHA A 135.5-1988, American National Standards Institute, New York, 1988.
10. *Medium-Density Fiberboard for Interior Use*, ANSI A208.2-1980, American National Standards Institute, New York, 1980.
11. *Hardboard Siding*, ANSI AHA A 135.6-1990, American National Standards Institute, New York, 1990.
12. *Standard Methods of Evaluating the Properties of Wood-Base Fiber and Particle Panel Materials*, ASTM D1037-93, American Society for Testing and Materials, Philadelphia, Pa.,
13. *Medium-Density Fiberboard (MDF)*, ANSI A208.2-1994, American National Standards Institute, New York, 1994.
14. *Particleboard*, ANSI A208.1-1993, American National Standards Institute, New York, 1993.
15. ASTM 1037-93, American Society for Testing and Materials, Philadelphia, Pa.,
16. *Performance Standard for Wood-Based Structural-Use Panels*, PS2-92, APA—The Engineered Wood Association, Tacoma, Wash., 1992.
17. *Standard Specification for Evaluation of Structural Lumber Composite Products*, ASTM D5456-93, American Society for Testing and Materials, Philadelphia, Pa., 1993.

General References

For more information, see publications of the following associations:
American Hardboard Association (AHA), 887-B Wilmette Road, Palatine, IL 60067, Tel. (708) 934-8800.
American National Standards Institute, 1430 Broadway, New York, NY 10018.
APA—The Engineered Wood Association, 7011 So. 19th St., Tacoma, WA 98411, Tel. (206) 565-6600.
The American Society for Testing and Materials (ASTM), 1916 Race St., Philadelphia, PA 19103–1187.
National Particleboard Association, 18928 Premiere Court, Gaithersburg, MD 20879, Tel. (301) 670–0604.
The Hardwood Plywood and Veneer Association (HPVA), 1825 Michael Faraday Drive, P.O. Box 2789, Reston, VA 22090, (703) 435–2900.

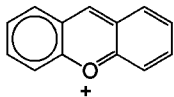
William F. Lehmann
Consultant

WOOD PULP.

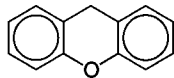
See PULP.

XANTHENE DYES

Xanthene dyes are those containing the xanthylium [261-23-4] (1a) or dibenzo- γ -pryan nucleus [92-83-1] (xanthene) (1b) as the chromophore with amino or hydroxy groups meta to the oxygen as the usual auxochromes. They are



(1a)

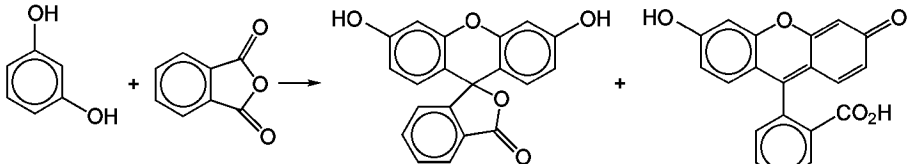


(1b)

important because of their brilliant hues, and shades between greenish yellows to dark violets and blues are obtainable, the most important being reds and pinks. They are generally very strong, with much higher oscillator strengths (and hence higher absorbances) than, for example, an anthraquinone dye of the same shade (see DYES, ANTHRAQUINONE). As a consequence of their rigid chromophoric nucleus, xanthenes are often fluorescent, which also adds to their strength and brightness, but, as is often the case with fluorescent dyes, they have lower lightfastness compared to other chromophores. Their use then is concentrated on those areas in which lightfastness is relatively unimportant compared to economy (eg, paper dyes) or where lightfastness can be achieved by modification by metallization. They are used for the direct dyeing of wool and silk and mordant dyeing of cotton. Paper, leather, woods, food, drugs, and cosmetics are dyed with xanthene dyes (see DYES, APPLICATION AND EVALUATION). Brilliant insoluble lakes are used in paints and varnishes. In addition, several new applications for xanthene dyes have emerged, for example in ink-jet printers, as markers in biological and medical research, and even as insecticides.

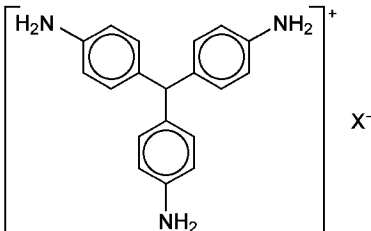
Xanthenes date from 1871 when von Bayer synthesized fluorescein (5) by the condensation of two moles of resorcinol with one mole of phthalic anhydride in the presence of concentrated sulfuric acid (1).

Depending on the reaction conditions, the product can be isolated in either the lactoid form A [2321-07-5] (2) or the quinonoid form B [56503-30-1] (3). These 9-phenylxanthenes are closely related structurally to the triphenyl methane dyes (4) and, like them, are cationic resonance hybrids.

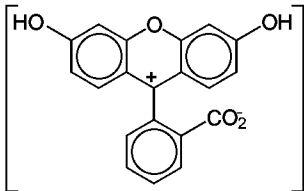


(2)

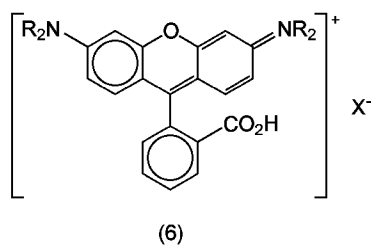
(3)



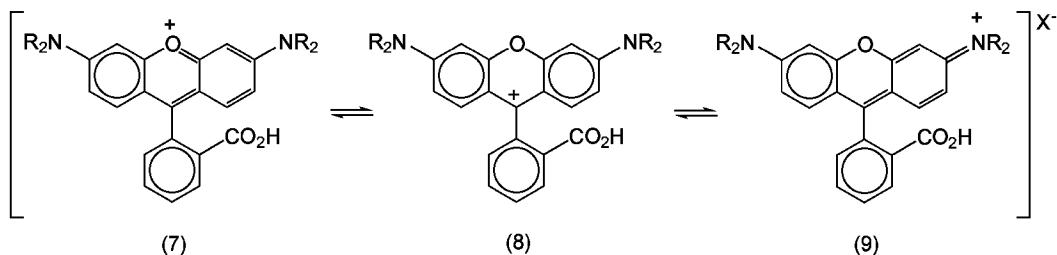
(4)



(5)



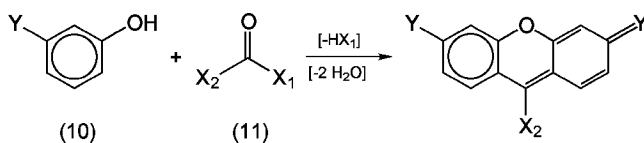
The main contributing resonance forms include the oxonium (7), carbonium (8), and ammonium structures (9):



For uniformity with the structures given in the *Colour Index* the ammonium radical (9) is used for the amino-substituted xanthenes and the keto form for the hydroxy derivatives. The xanthene dyes may be classified into two main groups: diphenylmethane derivatives, called pyronines, and triphenylmethane derivatives (eg, (4)), which are mainly phthaleins made from phthalic anhydride condensations. A third much smaller group of rosamines (9-phenylxanthenes) is prepared from substituted benzaldehydes. The phthaleins may be further subdivided into the following: fluoresceins (hydroxy-substituted); rhodamines (amino-substituted), eg, (6); and mixed hydroxy amino-substituted.

Most xanthene dyes are classified as basic dyes by their method of application; acid dyes can be produced by introduction of sulfonic acid groups. The fluoresceins, which contain carboxy and hydroxy substituents, are also acid dyes for coloration of silk. Some of the fluoresceins in which the carboxy group has been esterified, are soluble in alcohol or other organic solvents and can be classified as solvent dyes. Mordant dyes can be produced by introducing *o*-dihydroxy or salicylic acid groups (2), which when metallized can have very good lightfastness.

In general, the xanthenes are synthesized by the reaction of two moles of a nucleophilic *m*-substituted phenol (10) with an electrophilic carbonyl compound (11), the reaction occurring most readily with an acid catalyst at temperatures of 100–200°C.

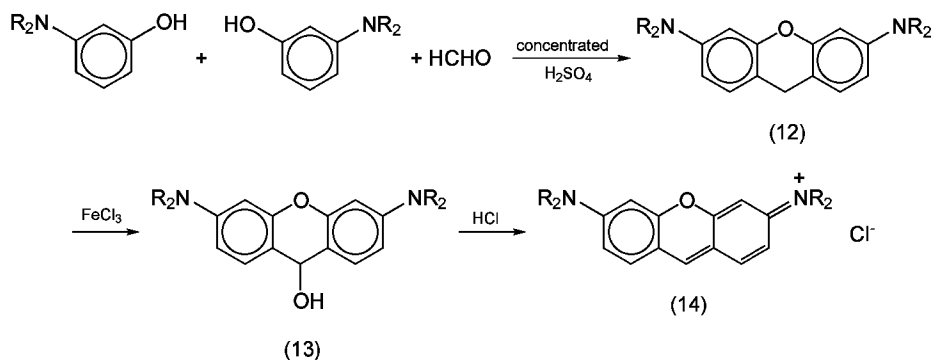


The physical properties of the xanthene type dye structure in general have been considered. For example, the aggregation phenomena of xanthene dyes has been reviewed (3), as has their photochemistry (4), electron transfer (5), triplet absorption spectra (6), and photodegradation (7). For the fluoresceins in particular, spectral properties and photochemistry have been reviewed (8), and the photochemistry of rhodamines has been investigated (9).

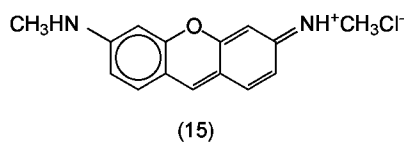
Recently, there have been many new uses for xanthene dyes reported. These include photo-activated insecticidal activity and pest-control (10,11), as anticancer agents (12), in hydrogen generation during the photolysis of water (13), as nonlinear optical (NLO) materials (14), or charge control agents in electrophotographic toners (laser printers and photocopiers) (15), and many reports as markers or biological stains (16–19). Rhodamines and rosamines in particular have been used in inks for ink-jet printers. There has also been a significant amount of new activity on the use of xanthenes, and in particular rhodamines, as laser dyes (20–22).

Diphenylmethane Derivatives

Pyronines. Pyronines are diphenylmethane derivatives synthesized by the condensation of *m*-dialkylaminophenols with formaldehyde, followed by oxidation of the xanthene derivative (12) to the corresponding xanthidrol (13) which in the presence of acid forms the dye (14). If R is methyl, the dye produced is

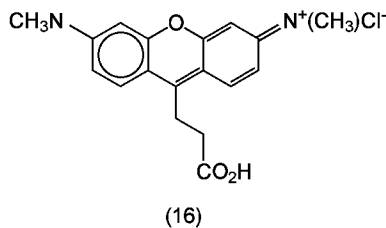


pyronine G (*CI 45005*); if R is ethyl, pyronine B (*CI 45010*) is obtained. If pyronine G is oxidized with potassium permanganate, two methyl groups are eliminated to give Acridine Red 3B (*CI 45000*) (15). Pyronine B and Acridine Red 3B



are used as biological stains and particularly in wet staining for the direct microscopic observations of living cells (23).

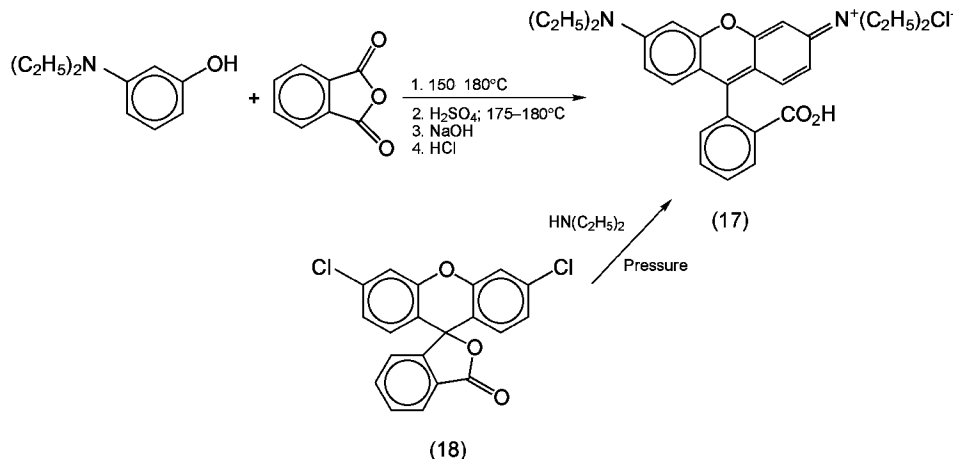
Succineins. Succineins are carboxyethyl-substituted pyronines made by substituting succinic anhydride for formaldehyde in the basic synthesis, as, for example, in Basic Red 11 [72968-14-0] (*CI 45050*) (Rhodamine S) (16).



Triphenylmethane Derivatives

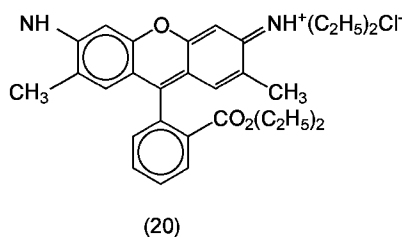
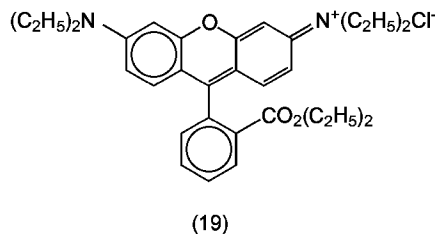
Amino-Derivatives.

Rhodamines. Rhodamines are commercially the most important aminoxanthenes. If phthalic anhydride is used in place of formaldehyde in the above condensation reaction with *m*-dialkylaminophenol, a triphenylmethane analogue, 9-phenylxanthene, is produced. Historically, these have been called rhodamines. Rhodamine B (Basic Violet 10, *CI* 45170) (17) is usually manufactured by the condensation of two moles of *m*-diethylaminophenol with phthalic anhydride (24). An alternative route is the reaction of diethylamine with fluorescein dichloride [630-88-6] (3,6-dichlorofluoran) (18) under pressure.



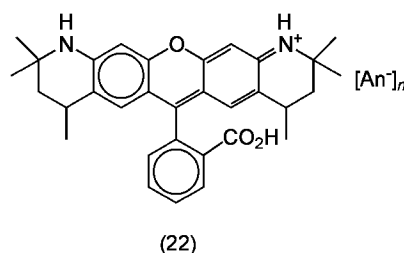
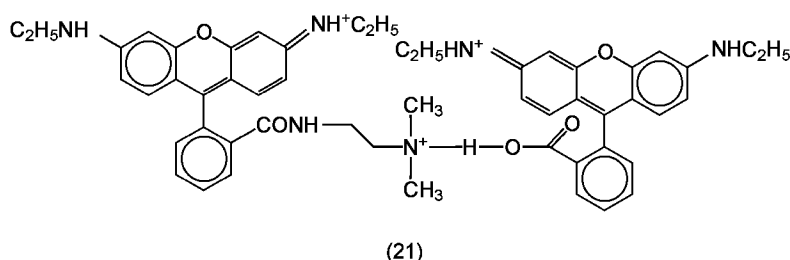
The free base of compound (17) is Rhodamine B base [509-34-2] (Solvent Red 49; *CI* 45170:1). The phosphotungstomolybdic acid salt of (17) is Pigment Violet 1 [1326-03-0] (*CI* 45170:2). Pigment Red 173 [12227-77-9] (*CI* 45170:3) is the corresponding aluminum salt.

Esterification of the carboxyl group also yields commercially useful dyes. If Rhodamine B is esterified with ethyl chloride or ethanol at 160–170°C under pressure, Basic Violet 11 [2390-63-8] (19) (*CI* 45175) forms. Another commercially important esterified aminoxanthene is Rhodamine 6G (Basic Red 1; *CI* 45160) [989-38-8] (20). This is manufactured by condensing 3-ethylamino-*p*-cresol with phthalic anhydride, then esterifying the product with ethanol and a mineral acid. The phosphotungstomolybdic acid salt of Rhodamine 6G is Pigment Red 81 [12224-98-5] (*CI* 45160:1). The copper ferrocyanide complex of structure (20) is Pigment Red 169 [122237-63-7] (*CI* 45160:2). Highly concentrated stable liquid forms of rhodamines can be prepared on a commercial scale by the reaction of the rhodamine base with dialkyl sulfate and a saturated aliphatic glycol at 100–160°C (25,26). These solutions are particularly suited for dyeing paper.

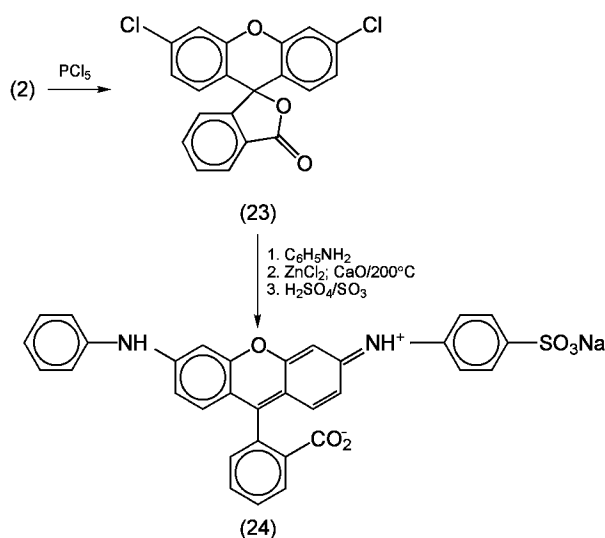


The rhodamines described thus far are basic rhodamines. They are used primarily for the dyeing of paper and the preparation of lakes for use as pigments. They are also used in the dyeing of silk and wool where brilliant shades with fluorescent effects are required, but where lightfastness is unimportant. Many new uses for rhodamine dyes have been reported. For example, when vacuum-sublimed onto a video disk, Rhodamine B loses its color to form a clear stable film which becomes permanently colored on exposure to uv light (27). This can be used in optical recording for computer storage or video recording (28). Fluorescent coloration of rigid or nonplasticized poly(vinyl chloride) (PVC) by the addition of selected rhodamines to the PVC resin before sheet or film formation has been reported (29). The addition of rhodamines to liquefied bis(hydroxyalkyl)aromatic dicarboxylic acid esters prior to their condensation to form polyesters to produce tinted polymer sheets is also used (30). Rhodamine 6G can be used in ink-jet printing where fluorescence under uv conditions is a desired property (31). Selected xanthenes, including fluorescein, Rhodamine B, and Rhodamine 6G, have been used as laser dyes (32).

Recently, several rhodamines have been used as magentas in ink-jet printing. These have been chosen primarily for their brilliant magenta shades, but with the disadvantage of very poor light- and waterfastness on ordinary copy papers. On specially coated color ink-jet papers the waterfastness improves, but the lightfastness is still much lower than comparable azo dyes, and, on this media, the difference in brightness between azo and xanthene chromophores is also much smaller. Attempts to improve the waterfastness of xanthene dyes for ink-jet have been made as in (21) (33,34). Improvements in lightfastness have been claimed by introducing branching and rings in the alkyl chains on the nitrogens and by using a polymeric counterion, (An)_n, eg, (22) (35).

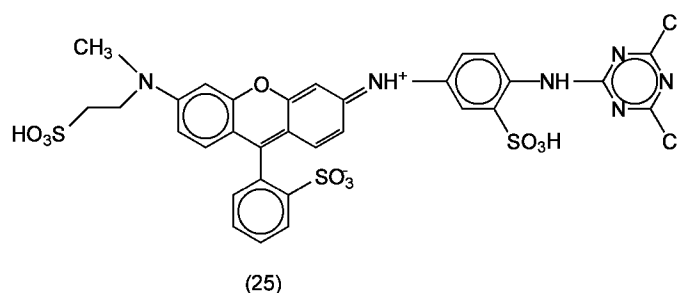


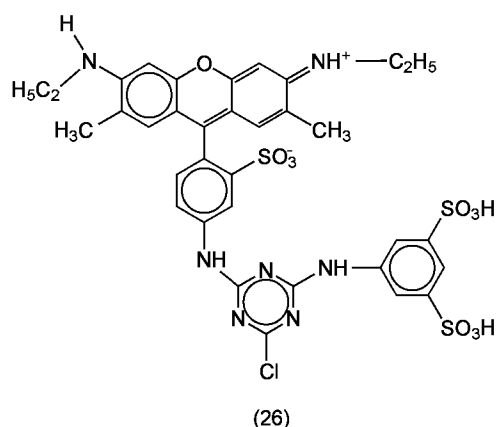
Acid rhodamines are made by the introduction of the sulfonic acid group to the aminoxanthene base. The preferred route is the reaction fluorescein (2) with phosphorous pentachloride to give 3,6-dichlorofluoran (fluorescein dichloride) (23), which is then condensed with a primary aromatic amine in the presence of zinc chloride and quicklime. This product is then sulfonated. For example, if compound (23) (fluorescein dichloride) is condensed with aniline and the product is sulfonated, Acid Violet 30 (*CI 45186*) (24) is produced.



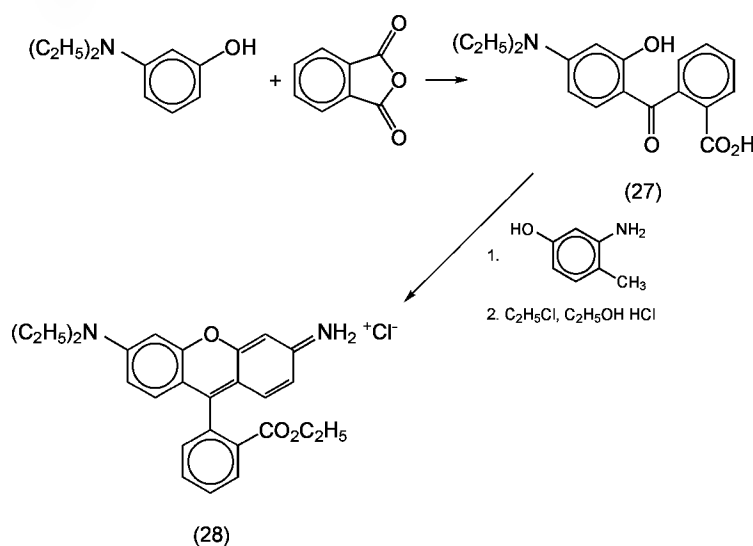
These acid rhodamines are usually used for silk and wool because they have level dyeing properties and show good fastness to alkali; however, they have poor lightfastness. An improved process for manufacturing 3,6-diaminosubstituted xanthenes is reaction of the inner salts of 3,6-dihalo-9-(2-sulfophenyl)xanthene-9-ols with a primary or secondary amine in stoichiometric amounts in the presence of an inorganic acid-binding agent or acid-binding tertiary aliphatic or tertiary nitrogen-containing heterocyclic amine (36).

Highly substituted acid rhodamines have been reported for fiber-reactive dye applications. For example, if 3,6-dichloro-9-(2-sulfophenyl)xanthene is first condensed with 1,4-phenylenediamine-3-sulfonic acid and then *N*-methyltaurine and then acylated with cyanuric chloride, compound (25) is produced. This sulfurein derivative dyes cellulosic fibers blue; shades have good washfastness and improved lightfastness (37). Another route to fiber-reactive xanthenes is exemplified by the condensation of 4-nitro-2-sulfobenzaldehyde and 3-*N*-ethylamino-4-methylphenol. The product is reduced and then reacts with an acylating agent formed by the reaction of 1-amino-3,5-disulfonic acid and cyanuric chloride to produce compound (26), which gives brilliant red shades with good washfastness and moderate fastness to light (38). In general then, it appears that the primary fading mechanisms for rhodamines is *N*-dealkylation, and substitution of the *N*-alkyl groups with *N*-aryl groups, ring fusions, or sterically hindered alkyls, which are less easily removed causes an increase in lightfastness.

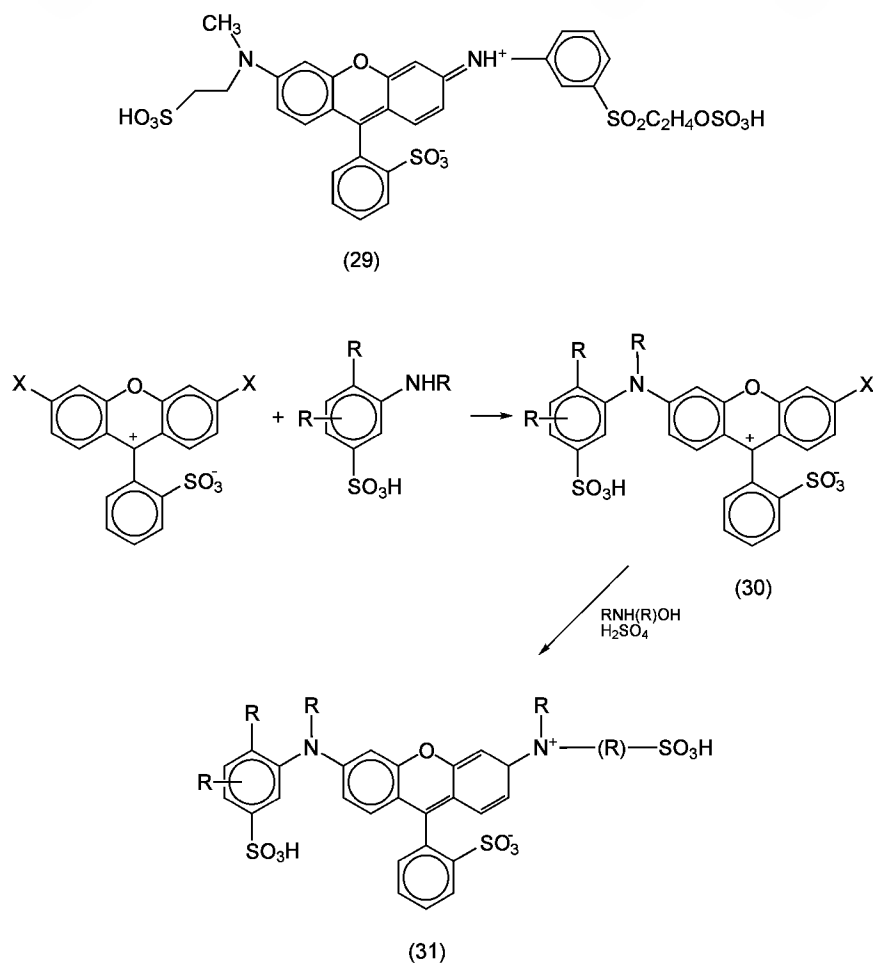




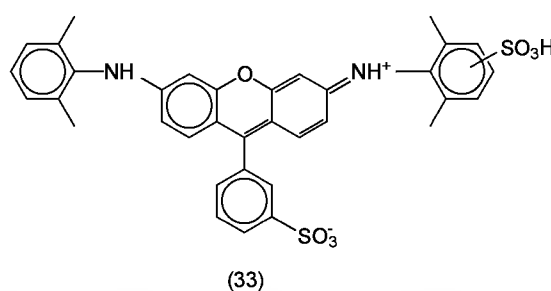
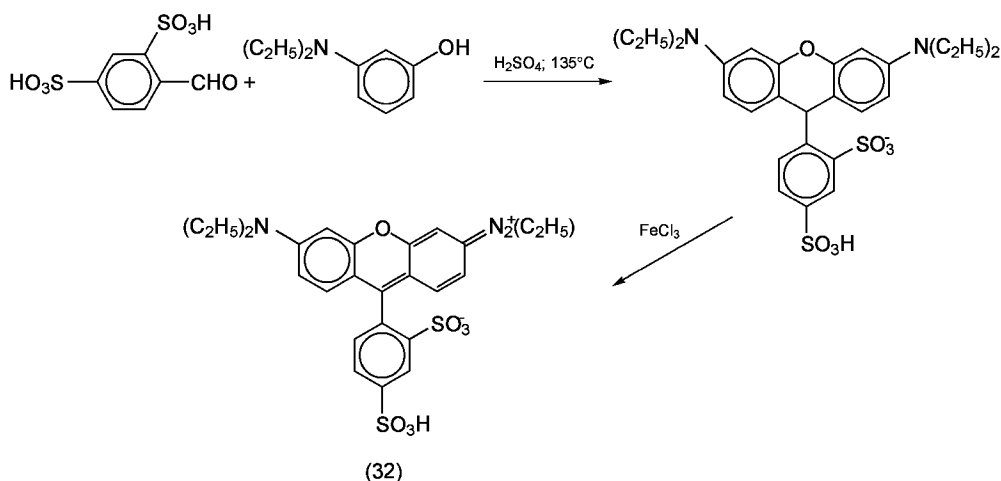
Unsymmetrical rhodamines can be prepared by the condensations of one mole of a *m*-aminophenol with phthalic anhydride to give an *o*-benzoyl benzoic acid (27) which is then further condensed with a different *m*-aminophenol to give the required product, Rhodamine 3GO (28) (2). A general route to asymmetrical acid xanthenes has been patented (39).



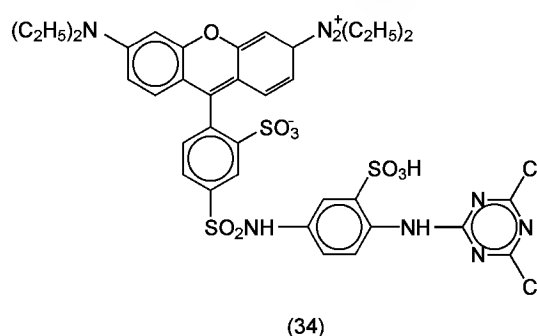
Reactive xanthene dyes with β -hydroxyethylsulfonyl groups, as exemplified by structure (29), provide brilliant shades and excellent washfastness on cotton (40). The sulfurein derivative (29) is synthesized by condensing 3-aminophenol- β -hydroxyethylsulfone with 3,6-dichloroxanthene-9-phenyl-2'-sulfonate at 90°C in *N*-methylpyrrolidone or dimethylformamide, then condensing with *N*-methyltaurine, and finally esterifying with chlorosulfonic acid (41). Condensation of 3,6-dihaloxanthene-9-phenyl-2'-sulfonic acid with an aromatic amine yields compound (30), a second condensation with an appropriately substituted aliphatic amine, usually with subsequent sulfonation, yields the acid rhodamine (31).



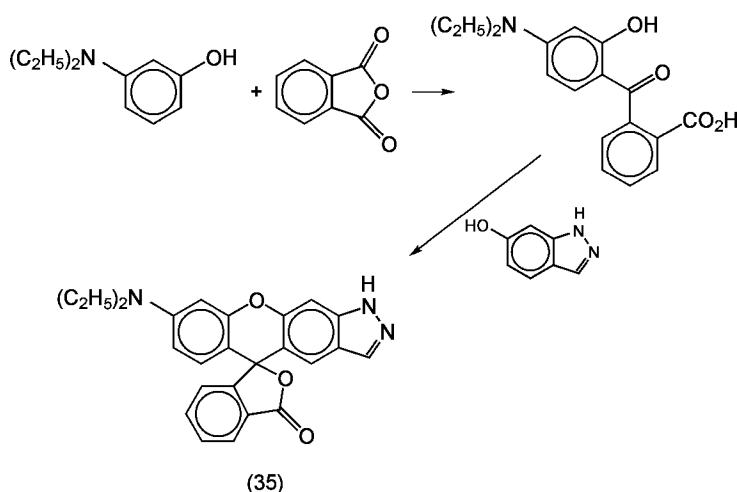
Rosamines. Rosamines are 9-phenylxanthene derivatives prepared from substituted benzaldehydes instead of phthalic anhydride. Condensing benzaldehyde-2,4-disulfonic acid with *m*-diethylaminophenol, dehydrating the product with sulfuric acid, and oxidizing with ferric chloride yields Sulforhodamine B [3520-42-1] (Acid Red 52; CI 45100) (32) which is the most important rosamine. Recently Acid Red 52 has been used in ink-jet printing as a magenta dye, sometimes in combination with other azo-magenta dyes to improve the lightfastness. The related compound Acid Red 289 (CI 45110) [12220-28-9] (33) can also be used and has slightly higher lightfastness. Both compounds have low waterfastness however, and cannot compete with the lightfastness achievable with an azo chromophore.



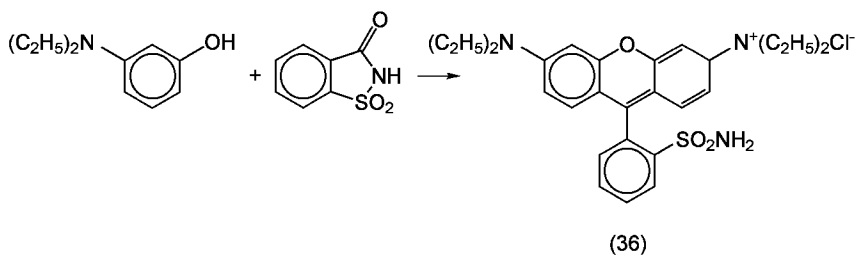
A series of fiber-reactive dyes have been made by the reaction of Sulforhodamine B with chlorosulfonic acid, an appropriately substituted diamine, and cyanuric chloride to yield dyes, eg, a Sulforhodamine B derivative (34), with good lightfastness (42).



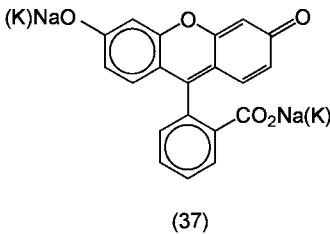
A group of aminoxanthenes, ie, pyrazoloxanthenes, is used as color formers in pressure or heat-sensitive imaging papers (43). These compounds are colorless, but, upon contact with acidic electron-accepting material, are converted to resonance forms that are lightly colored. An example is structure [58294-05-6] (35), which forms upon the condensation of *N,N*-diethyl-*m*-aminophenol with phthalic anhydride, followed by addition of 6-hydroxyindazole in 80% sulfuric acid (44).



Saccharein. Saccharein [6837-69-0] (36) is a basic dye prepared by the condensation of *m*-diethylaminophenol with saccharin at 165°C.

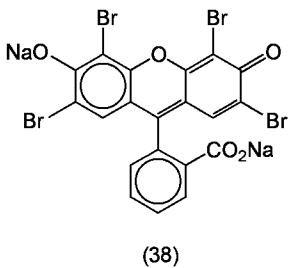


Hydroxyl Derivatives. The building block of most hydroxyl-substituted xanthenes, or fluorones, is fluorescein (2), (3). Fluoresceins can be made by the condensation of resorcinol with phthalic anhydride. Although fluorescein itself is no longer in general use as a textile dye, its intense green fluorescence, even at extremely high dilutions, makes it ideal for tracing water flows to detect leakage and as sea markers for downed aircraft and missing ships. However, Acid Red 388 is replacing fluorescein in such applications. The sodium or potassium salt of fluorescein, commonly called uranine [518-47-8] (CI 45350) (37), is still used for dyeing wool and silk brilliant yellow shades.

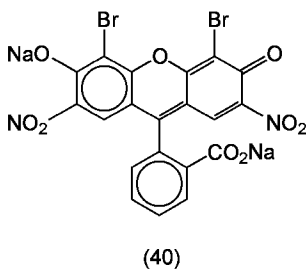
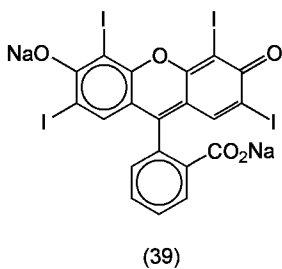


A modern use of uranine is in the manufacture of fluorescent laminates, eg, sheets, glass, and plastic films, that are transparent to electromagnetic waves and visible light rays (45). Such material might be used in windows, viewing partitions, and optical lenses.

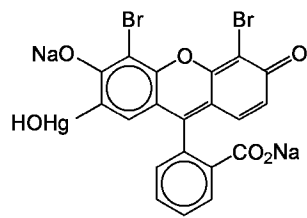
The principal use of fluorescein is as an intermediate for more highly substituted hydroxyxanthenes. When fluorescein is brominated in ethanolic solution and converted to the sodium salt with sodium chlorate, eosine [17372-87-1] (Acid Red 87; CI 45380) (38) forms. This has been shown to be the 2',4',5',7'-tetrabromo analogue. It is used for dyeing silk red with a brilliant yellow fluorescence, and for coloring inks, dyeing paper, and coloring cosmetics. Another use is as an indicator in the analytical determination of polymeric biguanide concentrations in aqueous solutions (see ANALYTICAL METHODS), which is important in controlling the growth of bacteria and algae in swimming pools (46) (see WATER, TREATMENT OF SWIMMING POOLS, SPAS AND HOT TUBS). The lead salt of eosine is Pigment Red 90 [1326-05-2] (CI 45380:1), the free acid is Solvent Red 43 [15086-94-9] (CI 45380:2), and the aluminum salt is Pigment Red 90:1 [17372-87-1] (CI 45380:3).



When fluorescein is reacted with iodine and potassium iodate in an ethanolic solution and converted to the sodium salt, the tetraiodo analogue erythrosine [16423-68-0] (Acid Red 51, CI 45430) (39) forms. This is used as a food coloring, a sensitizer in photographic plates, and in microscopical stains (see COLORANTS FOR FOOD, DRUGS, AND COSMETICS; PHOTOGRAPHY). Nitrated fluoresceins are used in dye applications. For example, dibrominations of fluorescein in aqueous sodium hydroxide followed by treatment with mixed sulfuric–nitric acid yields the dibromo–dinitro analogue saffrosine [548-24-3] (Acid Red 91; CI 45400) (40). Saffrosine is used to make fade-resistant electrophotographic sheet material by incorporating such nitro-substituted xanthenes in the photoconductive layer (47).

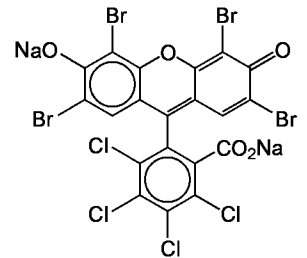


Treatment of this same 4',5'-dibromofluorescein intermediate with mercuric acetate and conversion to the disodium salt yields the hydroxymercuric analogue merbromin or mercurochrome [129-16-8] (41). It was once a widely used antiseptic, especially for skin disinfection, and was even administered internally. However, it has been replaced by more effective antibacterial agents.

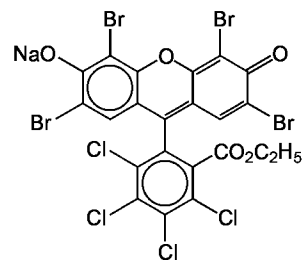


(41)

Another group of halogenated fluorescein dyes is prepared by condensing chloro derivatives of phthalic anhydride with resorcinol, followed by bromination or iodination. Thus Phloxine B [18472-87-2] (Acid Red 92, CI 45410) (42) is prepared by condensing tetrachlorophthalic anhydride with resorcinol followed by tetrabromination. Phloxine B undergoes ethylation to yield the yellowish red acid dye Cyanosine B [6441-80-1] (43).

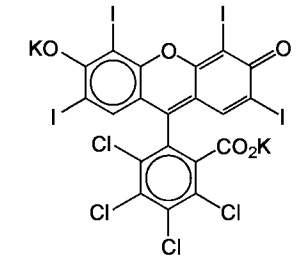


(42)



(43)

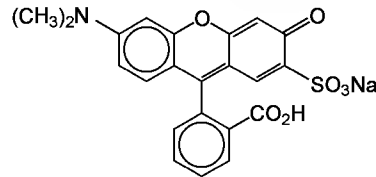
Another important polyhalogenated fluorescein dye is Rose Bengal [632-68-8] (Acid Red 94, CI 45440) (44). This is synthesized by condensation of resorcinol with tetrachlorophthalic anhydride, tetraiodination, and conversion to the potassium salt. A new use for Rose Bengal and other tetrabromo- or tetraiodofluoresceins involves their chemiluminescent reaction with ozone. Measuring the intensity of emitted light makes possible a quantitative determination of ozone concentration in the atmosphere (48) (see LUMINESCENT MATERIALS, CHEMILUMINESCENCE). Another recent use is as a nonsilver halide photographic system, particularly for use in making direct prints for microfilm enlargements (49).



(44)

Aminohydroxy Derivatives

Aminohydroxy-substituted xanthenes are of little commercial importance. They are synthesized by condensing one mole of *m*-dialkylaminophenol with phthalic anhydride, and then condensing that product with an appropriately substituted phenol. For example, Mordant Red 77 [6528-43-4] (CI 45300) (45) is prepared by condensing *m*-dimethylaminophenol with phthalic anhydride, and then condensing the product with 2,4-dihydroxybenzenesulfonic acid.

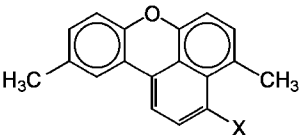


(45)

Miscellaneous Derivatives

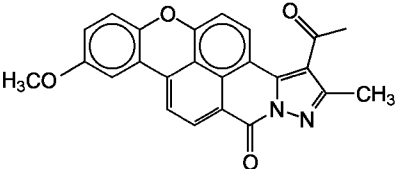
Two additional xanthene analogues are termed *fluorescent brighteners*. Condensation of two moles of *p*-cresol with one mole of phthalic anhydride yields 2',7'-dimethyl fluoran which is cyclized with 24° oleum and then reduced with zinc dust and ammonia under pressure to form Fluorescent Brightener 74 (CI 45550) (46). If the reduction is carried out with zinc dust and caustic soda in the presence of pyridine, followed by acetylation with acetic anhydride, Fluorescent Brightener 155 (CI 45555) (47) is obtained. These are used in the formulation of solid dielectric compositions for application in high voltage cables to prevent conductive treeing (50). The naphthalene analogue of fluorescein can be made by melting 1,6-dihydroxynaphthalene and phthalic

anhydride (51) and has been sold as Scheckfarbe AS, for security applications.

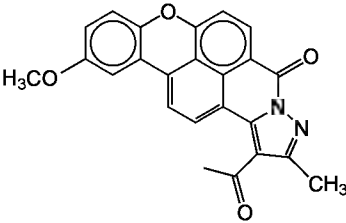


X = H (46)
X = O₂CCH₃ (47)

Another series of ring-closed xanthenes starting with benzoxanthene- and benzothioxanthenedicarboxylic acid hydrazides provides shades of bright yellow to red on cellulose acetate, polyamide and polyesters with excellent sublimation fastness and unusually good lightfastness. Because of their high fluorescence, they can be used for preparing daylight fluorescent pigments. For example, condensation of 8-methoxybenzo[*d,e*]xanthene-3,4-dicarboxylic acid hydrazide with acetylacetone in the presence of toluene sulfonic acid followed by cyclization of the hydrazone in *N*-methylimidizolidinone yields a mixture of isomers (48) and (49) (52).



(48)



(49)

A series of water-soluble fiber-reactive xanthene dyes has been prepared from the reaction of benzoxanthenedicarboxylic acid anhydride disulfonic acid with, for example, 3-aminophenyl-β-hydroxyethyl sulfone to yield dyes, with high brilliance and good fastness properties for dyeing of or printing on leather, wool, silk, or cellulosic fibers (53).

Economic Aspects

Since 1973, the U.S. International Trade Commission has reported the manufacture and sales of dyes by application class only. During the latter half of the 1970s and all through the 1980s, annual dye production in the United States, including xanthene dyes, changed very little. Statistics for the production of basic dyes include those products listed as cationic dyes, eg, cyanines, for dyeing polyacrylonitrile fibers and the classical triarylmethane dyes, eg, malachite green, for coloring paper and other applications. Furthermore, statistics for xanthene dyes are also hidden in the production figures for acid, solvent, mordant, and food dyes and organic pigments. Between 1975 and 1984, the production of basic dyes in the United States varied from (11–17 million lbs). However, from 1985–1990, production of basic dyes varied from (11–12.5 million lbs) with an increase in sales value from \$56 to \$73 million per annum. The production figures for xanthenes in 1980 are reproduced in Table 1.

Table 1. Economic Aspects of Xanthene Dyes

CI name	CAS Registry Number	CI Number	Common name	U.S. Imports (U.S. Production), t					U.S. suppliers
				1976	197	1978	1979	1980	
Acid Yellow 73	[2321-07-5]	45350	fluorescein uranine	0.45	0.7	1.6	1.7	2.0	American Cyanimid Co.; Leeben Color, Div. of Tricon Colors, Inc.; International Dyestuffs Corp.; Hilton-Davis Chemical Group of Sterling Drug, Inc. Atlantic Chemical Corp.; Carolina Color and Chemical Corp.; American Hoechst Corp.; Organic Chemical Corp.; San-doz Colours and Chemicals, Inc. Hilton-Davis Chemical Group of Sterling Drug, Inc. H. Kohnstann & Co.
Solvent Yellow 94	[518-47-8]	45350:1							
Acid Red 52	[3520-42-1]	45100	Sulforhoda-mi ne B	2.7	5.4	3.6	9.1	7.3	
Acid Red 87	[17372-87-1]	45380	eosine	0.45		0.7	3.2	2.7	
Pigment Red 90	[1326-05-2]	45380:1	Phloxine B	3.4	1.1	0.2	0.2	0.2	Hilton-Davis Chemical Group of Sterling Drug, Inc.; Sun Chemical Corp.
Solvent Red 43	[15086-94-9]	45380:2							
Acid Red 92	[18472-87-2]	45410	Rose Bengal		0.9	1.4	1.4	0.1	Hilton-Davis Chemical Group of Sterling Drug, Inc.
Solvent Red 48	[13473-26-2]	45410:1							
Pigment Red 174	[15876-58-1]	45410:2	Rose Bengal		0.9	1.4	1.4	0.1	Hilton-Davis Chemical Group of Sterling Drug, Inc.
Acid Red 94	[632-68-8]	45440							

Acid Red 289	[12220-28-9]			6.8			2.7	4.5	American Hoechst Corp. International Dyestuffs Corp.
Acid Violet 9	[6252-76-2]	45190		1.4	0.2	2.7	2.3	0.9	
Solvent Violet 10	[66225-66-9]								
Basic Red 1	[989-38-8]	45160	Rhodamine 6G	34	41	113	141	140	
Pigment Red 81	[12224-98-5]	45160:1		(263)	(233)	(254)	(227)	(186)	Atlantic Chemical Corp.; C. Lever Co.; International Dyestuffs Corp.; BASF Wyandotte Corp.; Dye Specialties, Inc.; C. Lever Co.; Sun Chemical Corp.; Mobay Chemical Corp.; CIBA-GEIGY Corp.; BASF Wyandotte Corp.
Basic Violet 10	[81-88-9]	45170	Rhodamine B	91	82	102	154	75	
Solvent Red 49	[509-34-2]	45170:1	Rhodamine B Base	6.8	11	21	2.7	14	
Pigment Violet 1	[1326-03-0]	45170:2			(64)	(62)	(120)	(104)	
Basic Violet 11	[2390-63-8]	45175	Fanal Red 6BM (1G)	29	14	22	27	16	Atlantic Chemical Corp.; American Cyanimid Co.; Buffalo Color Corp.; Sun Chemical Corp.; Leeben Color, Div. of Tricon Colors, Inc.; International Dyestuffs Corp.; BASF Wyandotte Corp.; Dye Specialties, Inc.; Mobay Chemical Corp.; American Cyanimid Co.; C. Lever Co.; Buffalo Color Corp.; International Dyestuffs Corp.; Dye Specialties, Inc.; CIBA-GEIGY-Corp.; Sun Chemical Corp.
Solvent Green 4	[81-37-8]	45550	Fluorescent Brightener 74	0.05	0.2	0.1			
Mordant Red 27	[6539-22-4]	45180	Chromoxane Brilliant Red				4.5	4.5	
Food Red 14	[16423-68-0]	45430	erythrosine						
									Warner Jenkinson Co.; Hilton-Davis Chemical Group of Sterling Drug, Inc.; Crompton and Knowles Corp.; Leeben Color, Division of Tricon Colors, Inc.

The rhodamines are economically the most important amino-substituted xanthene dyes. The total sales of Rhodamine B in the United States in 1980 were over \$10⁷. The total domestic market for fluorescein and uranine was estimated to be over \$0.5 × 10⁶ yr.

Health and Safety Factors, Toxicology

Xanthene dyes have not exhibited health or safety properties warranting special precautions; however, standard chemical labeling instructions are required. Toxicological properties of important dyes are listed in Table 2 (12).

Table 2. Toxicological Properties of Selected Xanthene Dyes^a

Compound	Structure	Property	Value, mg kg
xanthene	(1b)	LD ₅₀ (mouse), subcutaneous	690
fluorescein	(5)	LD _{Lo} (rat), intraperitoneal	600
		LD _{Lo} (mouse) ^b	600
		LD _{Lo} (rabbit), intravenous	300
		LD _{Lo} (guinea, pig) ^b	400
eosine	(30)	LD _{Lo} (rat), intraperitoneal	500
		LD _{Lo} (rat), subcutaneous	1,500
		TD _{Lo} (rat), subcutaneous	1,300
		LD ₅₀ (mouse), intravenous	550
		LD _{Lo} (rabbit), intravenous	300
		LD ₅₀ (rat), intraperitoneal	300
erythrosine	(31)	LD _{Lo} (rat), intravenous	200
		LD _{Lo} (rabbit), intravenous	200
		LD _{Lo} (mouse), oral	2,500
		LD ₅₀ (mouse), intravenous	370
		LD ₅₀ (mouse), intravenous	310
		TD _{Lo} (mouse), oral	39,600
Phloxine B	(34)	TD _{Lo} (mouse) ^b	66,000
		TD _{Lo} (rat) ^b	63,000
		LD _{Lo} (mouse), subcutaneous	20
		LD _{Lo} (rabbit), intravenous	15
		LD _{Lo} (mouse), intravenous	50

uranine	(29)	LD ₅₀ (rat), intraperitoneal	1,700
		TD _{Lo} (rat), subcutaneous	19
		LD ₅₀ (mouse), intraperitoneal	1,800
		LD ₅₀ (mouse), oral	4,700
		LD ₅₀ (rat), oral	6,700
		LD _{Lo} (guinea pig), intraperitoneal	1,000
Rhodamine B	(15)	LD _{Lo} (rat), oral	500
		TD _{Lo} (rat), subcutaneous	360
		LD ₅₀ (rat), intravenous	89,500
		LD _{Lo} (mouse), intraperitoneal	128
Rhodamine 6G	(18)	TD _{Lo} (rat), subcutaneous	100
		LD _{Lo} (mouse), intraperitoneal	2

^a Ref. 12.
^b Administration method unknown.

BIBLIOGRAPHY

"Xanthene Dyes" in *ECT* 1st ed., Vol. 15, pp. 136–149 by W. G Huey and S. K. Morse, General Aniline and Film Corp.; in *ECT* 2nd ed., Vol. 22, pp. 430–437 by F. F. Cesark, American Cyanamid Co.; in *ECT* 3rd ed., Vol. 24, pp. 662–677 by R. E. Farris, Sandoz Colors and Chemicals.

- A. Baeyer, *Chem. Ber.* **4**, 558, 662 (1871).
- H. A. Lubs, ed., *The Chemistry of Synthetic Dyes and Pigments*, American Chemical Society Monograph Series, Reinhold Publishing Corp., New York, 1955.
- O. Valdes-Aguilera and D. C. Neckers, *Acc. Chem. Res.* **22**(5), 171–177 (1989).
- O. Valdes-Aguilera and D. C. Neckers, *Adv. Photochem.* **18**, 315–394 (1993).
- A. K. Chibisov and G. V. Zakharova, *Usp. Nauchn. Fotogr.* **25**, 114–136 (1989).
- I. Cammichael and G. L. Hug, *J. Phys. Chem. Ref. Data* **15**(1), 1–250 (1986).
- Y. Y. Marchant, *ACS Symp. Ser.* **339**, 168–175 (1987).
- D. C. Neckers, *J. Photochem. Photobiol. A*, **471**(1), 1–29 (1989).
- A. F. Lopez, A. T. Lopez, L. E. Gil, and A. I. Lopez, *Appl. Fluoresc. Technol.* **2**(5), 8–12 (1990).
- J. R. Heitz, *ACS Symp. Ser.* **616**, 1–16 (1995).
- J. R. Heitz, in J. R. Coats, ed., *Insecticide Mode Action*, Academic Press, New York, 1982, pp. 429–457.
- D. V. Sweet, *Registry of Toxic Effects of Chemical Substances*, 1986 ed., U.S. Dept. of Health and Human Services, National Institute for Occupational Safety and Health, Cincinnati, Ohio.
- X. Zhang and T. Shen, *Hanxue Tongbao*, (6), 8–14 (1995).
- S. Speiser and F. L. Chisena, *Spec. Pub. R. Chem. Soc.* **69**, 211–216 (1989).
- Jpn. Pat. 6227850** (Dec. 3, 1987), N. Nakayama, T. Isoda, Y. Watanabe, M. Aoki, and co-workers (to Ricoh Co.).
- Brit. Pat. 2283744** (May 17, 1995), R. P. Haugland, M. N. Malekzadeh, and Y. Zhang (to Molecular Probes Inc.).
- X. He, X. Gu, G. Zhao, and G. Dai *Wuli Huaxue Xuebao*, **11**(6), 504–508 (1995).
- Brit. Pat. 2277097** (Oct. 19, 1994), M. A. Kuhn and R. P. Haughland (to Molecular Probes Inc.).
- M. Torneiro and W. C. Still, *J. Am. Chem. Soc.* **117**(21), 5887–5888 (1995).
- E. Reisfeld, *J. Phys. IV*, **4** (1994).
- U.S. Pat. 5,111,472** (May 5, 1992), P. R. Hammond and J. F. Feeman (to U.S. Dept. of Energy).
- U.S. Pat. 4,945,176** (July 31, 1990), P. R. Hammond and J. F. Feeman (to U.S. Dept. of Energy).
- U.S. Pat. 3,961,039** (June 1, 1976), R. Sternheimer.
- BIOS Report 959*, British Intelligence Objectives Subcommittees, U.K., 1946, pp. 8 and 15.
- U.S. Pat. 3,849,065** (Nov. 19, 1974), K. Schmeidl (to BASF).
- U.S. Pat. 3,767,358** (Oct. 23, 1973), H. I. Stryker (to DuPont).
- U.S. Pat. 3,690,889** (Sept. 12, 1972), S. E. Harrison and R. Drake (to RCA Corp.).
- U.S. Pat. 3,767,408** (Oct. 23, 1973), S. E. Harrison and J. E. Goldmacher (to RCA Corporation).
- U.S. Pat. 3,796** (Mar. 12, 1974), R. T. Hickcox (to Herculea, Inc.).
- U.S. Pat. 3,644,270** (Feb. 22, 1972), G. D. V. Aliaveedam (to DuPont).
- Brit. Pat. 1,494,768** (Dec. 14, 1977), D. M. Zabiak and K. S. Hwang (to A. B. Dick Co.).
- U.S. Pat. 3,541,470** (Nov. 17, 1970), J. R. Lankard and P. P. Sorokin (to IBM Corp.).
- U.S. Pat. 4,935,059** (June 23, 1988), U. Mayer and A. Oberdinner (to BASF).
- U.S. Pat. 4,647,675** (July 12, 1984), U. Mayer and A. Oberdinner (to BASF).
- U.S. Pat. 5,410,053** (May 11, 1992), B. Albert, W. Denziger, E. Kahn, and C. Kraeh (to BASF).
- Brit. Pat. 1,503,380** (Mar. 8, 1978) (to Hoechst AG).
- U.S. Pat. 3,956,300** (May 11, 1976), P. W. Austin, A. T. Costello, and A. Crabtree (to ICI Ltd.).
- Brit. Pat. 1,377,695** (Dec. 18, 1974), P. W. Austin and A. T. Costello (to ICI Ltd.).
- Brit. Pat. 1,586,820** (Mar. 25, 1981) (to Hoechst AG).
- Brit. Pat. 1,471,452** (Apr. 27, 1977) (to Hoechst AG); **Brit. Pat. 1,471,453** (Apr. 27, 1977) (to Hoechst AG).
- U.S. Pat. 3,772,335** (Nov. 13, 1973), F. Meininger and F. Kohlhass (to Hoechst AG).
- U.S. Pat. 3,883,529** (May 13, 1975), P. W. Austin (to ICI Ltd.).
- U.S. Pat. 3,988,492** (Oct. 26, 1976), S. M. Spatz (to Mead Corp.).
- U.S. Pat. 3,929,825** (Dec. 30, 1975), S. M. Spatz (to Mead Corp.).
- Brit. Pat. 1,429,597** (Mar. 24, 1978), G. O. Okikiolu (to Okikiolu Scientific and Industrial Organization).
- Brit. Pat. 1,533,255** (Nov. 22, 1978), G. G. Barraclough, D. Myles, D. A. Reilly, and P. Tomlinson (to ICI Ltd.).
- U.S. Pat. 3,951,655** (Apr. 20, 1976), B. Schoustra and H. Roncken (to Oce-van der Grinten NV, Netherlands).
- U.S. Pat. 3,975,159** (Aug. 17, 1976), S. Van Heusden (to U.S. Philips Corp.).
- U.S. Pat. 3,615,566** (Oct. 26, 1971), I. D. Robinson (to Arthur D. Little, Inc.).
- U.S. Pat. 4,216,101** (Aug. 5, 1980), H. J. Davis (to Canada Wire and Cable Ltd.).
- K. Ventkataraman, ed., *The Chemistry of Synthetic Dyes*, Vol. 2, Academic Press, Inc., New York, 1952.

52. U.S. Pat. 3,853,884 (Dec. 10, 1974), H. Troster (to Hoechst AG).
53. U.S. Pat. 3,888,862 (June 10, 1975), F. Meininger and F. Kohlhas (to Hoechst AG).

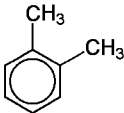
General References

D. R. Warning and G. Hallas, ed., *The Chemistry and Applications of Dyes*, Plenum Publishing Corp., New York, 1990.
P. Gregory, *High-Technology Applications of Organic Colorants*, Plenum Publishing Corp., New York, 1990.
G. Booth, *The Manufacture of Organic Colorants and Intermediates*, The Society of Dyers and Colourists, Bradford, U.K., 1988.
P. F. Gordon and P. Gregory, *Organic Chemistry in Colour*, Springer Verlag, Berlin, 1983.
G. Hallas, in J. Shore, *Colorants and Auxiliaries*, Vol. 1, Society of Dyers and Colourists, 1990, pp. 279–294.
E. N. Abrahart, *Dyes and their Intermediates*, 2nd ed., Chemical Publishing, New York, 1977.
R. L. M. Allen, *Color Chemistry*, Appleton-Century-Crofts, New York, 1971.
K. Ventkataraman, ed., *The Chemistry of Synthetic Dyes*, Vol. 2, Academic Press Inc., New York, 1952.
Color Index, 3rd ed., The Society of Dyers and Colourists, Bradford, U.K. and the American Associations of Textile Chemists and Colorists, N.C., Vols. 1–6, 1971.
P. Rys and H. Zollinger, *Fundamentals of the Chemistry and Applications of Dyes*, Wiley-Interscience, New York, 1972.
H. A. Lubs, ed., *The Chemistry of Synthetic Dyes and Pigments*, American Chemical Society Monograph Series, Reinhold Publishing Corp., New York, 1955.
K. Venkataraman ed., *The Analytical Chemistry of Synthetic Dyes*, Wiley-Interscience, New York, 1977.
J. Fabian and H. Hartmann, *Light Absorption of Organic Colorants*, Springer Verlag, Berlin, 1980.
T. E. Furia, ed., *Handbook of Food Additives*, 2nd ed., CRC Press, Cleveland, Ohio, 1972.
D. M. Marmion, *Handbook of U.S. Colorants for Food, Drugs, and Cosmetics*, John Wiley & Sons, Inc., New York, 1979.
E. Gurr, *Synthetic Dyes in Biology, Medicine and Chemistry*, Academic Press, London, 1971.
R. Raue, *Rev. Prog. Coloration*, **14**, 187 (1984).

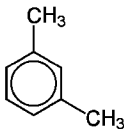
Paul Wight
Zeneca Specialties

XYLENES AND ETHYLBENZENE

Xylenes and ethylbenzene [100-41-4] (EB) are C₈ aromatic isomers having the molecular formula C₈H₁₀. The xylenes consist of three isomers; *o*-xylene [95-47-6] (OX), *m*-xylene [108-38-3] (MX), and *p*-xylene [106-42-3] (PX). These differ in the positions of the two methyl groups on the benzene ring. The molecular structures are shown below.



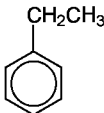
o-Xylene (OX)



m-Xylene (MX)



p-Xylene (PX)



Ethylbenzene (EB)

Sources and Uses

The term *mixed xylenes* describes a mixture containing the three xylene isomers and usually EB. Commercial sources of mixed xylenes include catalytic reformat, pyrolysis gasoline, toluene disproportionation product, and coke-oven light oil. Ethylbenzene is present in all of these sources except toluene disproportionation product. Catalytic reformat is the product obtained from catalytic reforming processes. In catalytic reforming, a low octane naphtha cut (typically a straight run or hydrocracked naphtha) is converted into high octane aromatics, including, benzene, toluene, and mixed xylenes (see BTX PROCESSING). Aromatics are separated from the reformat using a solvent such as diethylene glycol or sulfolane and then stripped from the solvent. Distillation is then used to separate the BTX into its components. The amount of xylenes contained in the catalytic reformat depends on the fraction and type of crude oil, the reformer operating conditions, and the catalyst used. The amount of xylenes produced can vary widely, typically ranging from 18 to 33 vol % of the reformat. In the United States, only about 12% of the xylenes produced via catalytic reforming is actually recovered for use as petrochemicals. The unrecovered reformat xylenes are used in the gasoline pool.

Pyrolysis gasoline is a by-product of the steam cracking of hydrocarbon feeds in ethylene crackers (see ETHYLENE). Pyrolysis gasoline typically contains about 50–70 wt % aromatics, of which roughly 50% is benzene, 30% is toluene, and 20% is mixed xylenes (which includes EB).

Coke oven light oil is a by-product of the manufacture of coke for the steel industry. When coal is subjected to high temperature carbonization, it yields 16–25 liters tonne of light oil that contains 3–6 vol % of mixed xylenes.

Toluene disproportionation (TDP) is a catalytic process in which 2 moles of toluene are converted to 1 mole of xylene and 1 mole of benzene; this process is discussed in greater detail herein. Although the mixed xylenes from TDP are generally more costly to produce than those from catalytic reformat or pyrolysis gasoline, their principal advantage is that they are very pure and contain essentially no EB.

A breakdown of the mixed xylene supply sources in the United States is summarized in Table 1 (1). As shown in Table 1, the primary source of xylenes in the United States is catalytic reformat. In 1992, over 90% of the isolated xylenes in the United States were derived from this source. Approximately 9% of the recovered xylenes is produced via toluene disproportionation (TDP). In the United States, only negligible amounts of the xylenes are recovered from pyrolysis gasoline and coke oven light oil. In other parts of the world, pyrolysis gasoline is a more important source of xylenes.

Table 1. U.S. Supply Source for Mixed Xylenes^a

	Xylenes, 10 ³ t
--	----------------------------

Sources	Contained	Recovered
catalytic reformat	31,603	3,766
pyrolysis gasoline	395	16
toluene disproportionation	362	362
coke oven light oil	3	3
<i>Total</i>	<i>32,363</i>	<i>4,147</i>

^a Ref. 1.

The mixed xylenes that are recovered from these sources are used as follows: 50–60° to make PX, 10–15° to make OX, 10–25° returned back to gasoline blending, and only 1° to make MX. The relative lack of end uses for MX is unfortunate, because these supply sources typically contain twice as much MX as PX or OX. The majority of the MX in these sources is isomerized to PX and OX, as discussed herein. The purified xylenes are used to synthesize plasticizers and polyester fibers, photographic films, and beverage bottles. PX is first oxidized to terephthalic acid or dimethyl terephthalate before being converted into polyesters. OX is oxidized to phthalic anhydride before being converted into plasticizers. MX is oxidized to isophthalic acid, which is used to make polyesters.

Physical Properties

Because of their similar molecular structures, the three xylenes and EB exhibit many similar properties (see Table 2). The very close boiling points of these compounds makes it difficult to separate them from each other by conventional distillation. OX is the easiest to distill from a mixture because of the 5°C difference in boiling point between it and the next closest boiling isomer, MX. This distillation is practiced commercially using one or two columns having a total of about 150 trays and a high reflux ratio. EB can also be separated from the mixture by distillation. However, this superfractionation requires several columns having a total of more than 300 theoretical trays. This method is highly energy-intensive compared to the production of EB via alkylation of benzene with ethylene. In the United States, Philbro Energy has operated a superfractionating facility, but this unit is currently shut down. Virtually no EB is currently made in the United States by superfractionation.

Table 2. Physical Properties for C₈ Aromatic Compounds^a

Property	<i>p</i> -Xylene	<i>m</i> -Xylene	<i>o</i> -Xylene	Ethylbenzene
molecular weight	106.167	106.167	106.167	106.167
density at 25°C, g cm ³	0.8610	0.8642	0.8802	0.8671
boiling point, °C	138.37	139.12	144.41	136.19
freezing point, °C	13.263	47.872	25.182	94.975
refractive index at 25°C	1.4958	1.4971	1.5054	1.4959
surface tension ^b , mN m(dyn cm)	28.27	31.23	32.5	31.50
dielectric constant at 25°C	2.27	2.367	2.568	2.412
dipole moment of liquid ^c C·m	0	0.30	0.51	0.36
critical properties ^b				
critical density, mmol cm ³	2.64	2.66	2.71	2.67
critical volume, cm ³ mol	379.0	376.0	369.0	374.0
critical pressure, MPa ^d	3.511	3.535	3.730	3.701
critical temperature, °C	343.05	343.90	357.15	343.05
thermodynamic properties ^e				
<i>C</i> _s at 25°C, J (mol·K) ^f	181.66	183.44	188.07	185.96
<i>S</i> _s at 25°C, J (mol·K) ^f	247.36	253.25	246.61	255.19
<i>H</i> _o – <i>H</i> _o at 25°C, J mol ^f	44.641	40.616	42.382	40.219
(<i>G</i> _s – <i>H</i> _o <i>T</i>) at 25°C, J (mol·K) ^f	97.633	117.03	104.46	120.29
heats of transition, J (mol·K) ^f				
vaporization at 25°C	42.036	42.036	43.413	42.226
formation at 25°C	24.43	25.418	–24.439	12.456
vapor pressure, Antoine equation ^g				
A	6.1155	6.1349	6.1239	6.0821
B	1453.430	1462.266	1474.679	1424.255
C	215.307	215.105	213.686	213.206

^a Convenient graphic representation of the change in property with temperature are given in Refs. 2 and 3.

^b Ref. 4.

^c To convert C·m to D, divide by 3.336 × 10^{–30}.

^d To convert MPa to psi, multiply by 145.

^e Refs. 5 and 6.

^f To convert J to cal, divide by 4.184.

^g log*P*_{kPa} = A – B / (C + t); (log*P*_{mmHg} = log*P*_{kPa} + log7.50).

PX is not separated via distillation because its boiling point is too close to that of MX. Instead, the differences in freezing points and adsorption characteristics are exploited commercially, as described in detail herein.

Since xylenes are important components of gasoline, their combustion and octane characteristics are of interest. The critical compression ratios are 14.2, 1.36, and 9.6 for PX, MX, and OX, respectively. The research octane numbers are 116.4, 117.5, 107.4, and 113 for PX, MX, OX, and EB, respectively (7). The motor octane numbers are 110.0, 111.5, 100, and 105 for PX, MX, OX, and EB respectively.

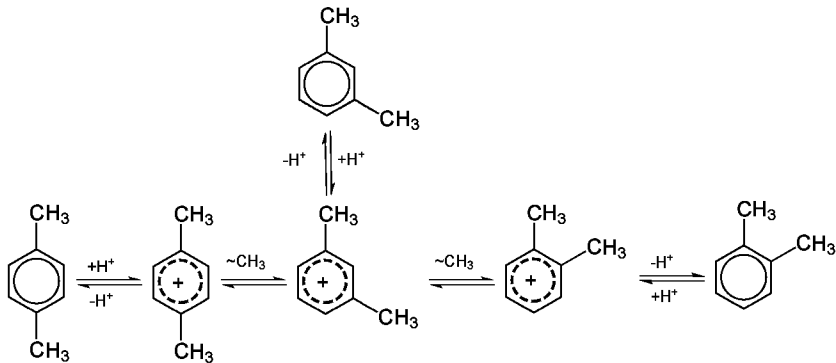
Chemical Properties

Chemical reactions that the xylenes participate in include (1) migration of the methyl groups, (2) reaction of the methyl groups, (3) reaction of the aromatic ring, and (4) complex formation.

Migration of the Methyl Groups. Reactions that involve migration of the methyl groups include isomerization, disproportionation, and

dealkylation. The interconversion of the three xylene isomers via isomerization is catalyzed by acids. The acids can be liquids or solids. One example of an acidic liquid-phase system is hydrogen flouride–boron trifluoride (8). At low boron trifluoride concentrations, the xylenes isomerize to near equilibrium levels. At high boron trifluoride concentrations, a complex containing MX–hydrogen fluoride–boron trifluoride is formed which can be decomposed to produce high purity MX. Two other acidic liquid-phase systems that can isomerize the xylenes are hydrogen bromide in toluene and aluminum bromide in toluene (9). Examples of solid acids include aluminum-based materials and zeolites.

The mechanism of these reactions involves the rapid and reversible addition of a proton to the aromatic ring, followed by 1,2-intramolecular methyl shifts (10):



As shown in Figure 1, the equilibrium concentration is affected slightly by temperature (11). The actual concentration is affected by the reaction rate and the initial concentration of each isomer. Deviations beyond equilibrium can be achieved when zeolites are used, owing to shape selectivity (see MOLECULAR SIEVES). The thermal isomerization of the three xylenes has been studied at 1000°C (12). Side reactions predominated, and only a small percentage of xylenes was interconverted.

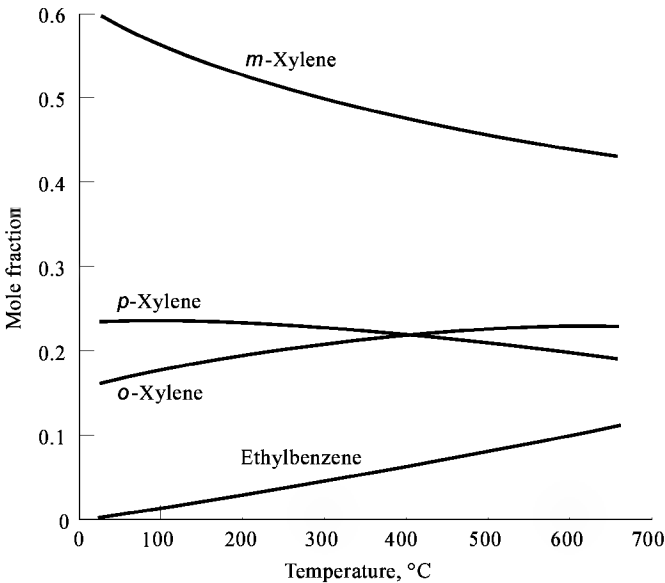
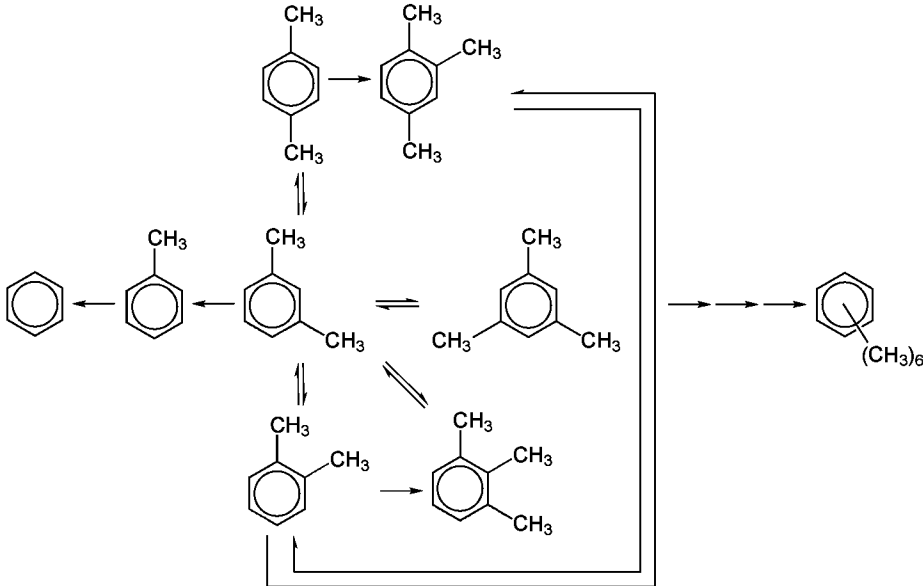


Fig. 1. Equilibrium concentrations for C₈ aromatic compounds (11).

Transalkylation is also catalyzed by acids, but requires more severe conditions than isomerization. As shown below, the methyl migration is intermolecular and ultimately produces a mixture of aromatic compounds ranging from benzene to hexamethylbenzene. The overall equilibrium constants for all possible methylbenzenes have been determined experimentally and calculated theoretically (Fig. 2 and Table 3).



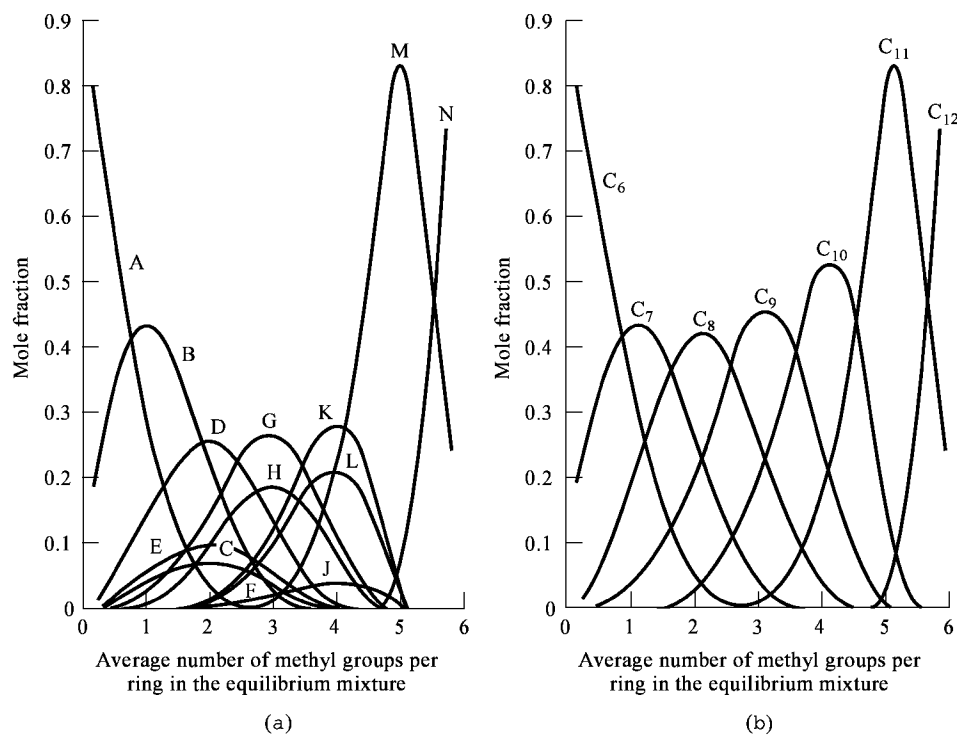


Fig. 2. Equilibrium concentrations of methyl benzenes in ideal gas state at 25°C (13). A – benzene; B – toluene; C – OX; D – MX; E – PX; F – hemimellitene; G – pseudocumene; H – mesitylene; J – prehnitene; K – isodurene; L – durene; M – pentamethylbenzene; N – hexamethylbenzene.

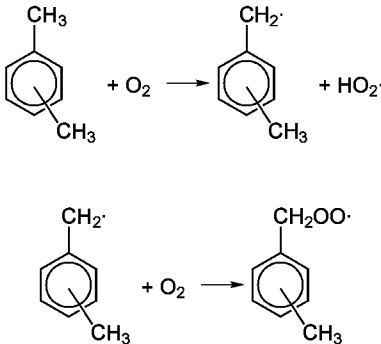
Table 3. Overall Equilibrium Constants of Methylbenzenes^a

Hydrocarbon	$K_{eq}^{b,c}$			
	Reference 15	Reference 16	Reference 17	HCI ^d and inductive model ^a
benzene		0.09	2×10^{-4}	$1-2 \times 10^{-4}$
toluene	~0.01	0.63	0.25	0.11–0.13
OX	2	1.1		1.8
MX	20	26	300	90–110
PX	1	1	1	1
pseudocumene	40	63		110–140
hemimellitene	40	69		200–310
durene	120	140		510–810
prehnitene	170	400		960–1700
mesitylene	2800	1.3×10^4	2×10^5	$2-3.6 \times 10^4$
isodurene	5600	1.6×10^4		$0.6-1 \times 10^5$
pentamethylbenzene	8700	2.9×10^4		$1.1-2.5 \times 10^5$
hexamethylbenzene	8.0×10^4	9.7×10^4	1×10^7	$0.7-1.5 \times 10^6$

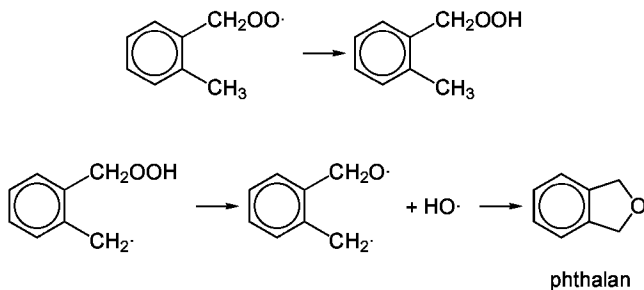
^a Ref. 14.
^b Compared to PX.
^c Column headings indicate sources of various test results.
^d HCI – hyperconjugative.

Reactions of the Methyl Groups. These reactions include oxidation, polycondensation, and amoxidation. PX can be oxidized to both terephthalic acid and dimethyl terephthalate, which are then condensed with ethylene glycol to form polyesters. Oxidation of OX yields phthalic anhydride, which is used in the production of esters. These are used as plasticizers for synthetic polymers. MX is oxidized to isophthalic acid, which is also converted to esters and eventually used in plasticizers and resins (see PHthalic ACIDS AND OTHER BENZENEPOLYCARBOXYLIC ACIDS).

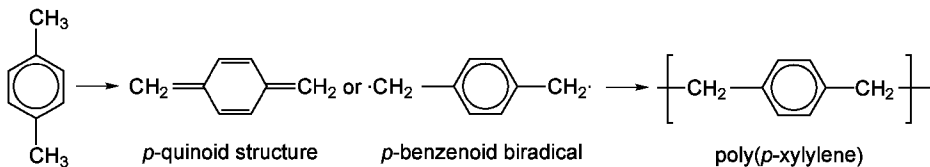
In a study of the slow combustion of the three xylenes, it was observed that OX is much more reactive toward oxygen than MX and PX (18). Under identical conditions, OX was approximately ten times as reactive as its isomers. It was proposed that the initial steps in the mechanisms of formation of each isomer are the same.



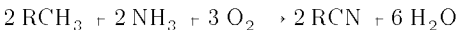
In the oxidation of PX and MX, formaldehyde is a degenerate branching intermediate, whereas phthalan is formed from OX:



PX forms *p*-xylylene when heated above 1200°C. The structure of *p*-xylylene is represented by a *p*-quinoid structure or as a *p*-benzenoid biradical. Condensation yields poly(*p*-xylylene) (19–22) (see XYLENE POLYMERS).

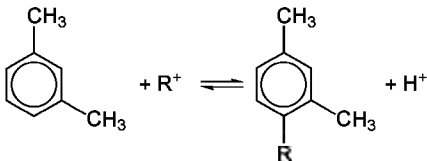


The methyl groups on xylenes can undergo ammoxidation, reaction with ammonia and oxygen (23).



The Showa Denka Company practices this reaction with a PX–MX mixture (24), whereas Mitsubishi Gas Chemical Company uses high purity MX first to form the dicyanide (25). In both processes, hydrogenation to the diamine follows. *m*-Xylenediamine is reacted with phosgene to give *m*-xylene diisocyanate, which is used in urethane resins (26–28).

Reactions of the Aromatic Ring. The reactions of the aromatic ring of the C₈ aromatic isomers are generally electrophilic substitution reactions. All of



the classical electrophilic substitution reactions are possible (see FRIEDEL CRAFTS REACTIONS), but in most instances they are of little practical significance. The relative nuclear chlorination rates of polymethylbenzenes have been studied (29,30). The higher the degree of substitution, the higher is the rate of chlorination.

As in most electrophilic reactions, the ability to stabilize the positive charge generated by the initial addition strongly affects the relative rates. MX reacts faster than OX and PX because both methyl groups work in conjunction to stabilize the charge on the next-but-one carbon. Sulfonation was, at one time, used to separate MX from the other C₈ aromatic isomers. MX reacts most rapidly to form the sulfonic acid which remains in the aqueous phase. The sulfonation reaction is reversible, and MX can be regenerated.

Hydrogenation of the aromatic ring to form naphthenic compounds has been proposed as a route to facilitate the separation of the C₈ aromatic isomers (31). The spread in boiling points of the naphthenic compounds is 12°C vs a spread of 8°C for the aromatic compounds. However, the cycloparaffinic products obtained from OX and EB boil only 3°C apart, impeding the separation.

Complex Formation. All four C₈ aromatic isomers have a strong tendency to form several different types of complexes. Complexes with electrophilic agents are utilized in xylene separation. The formation of the HF–BF₃–MX complex is the basis of the Mitsubishi Gas–Chemical Company (MGCC) commercial process for MX recovery, discussed herein. Equimolar complexes of MX and HBr (mp –77°C) and EB and HBr (mp –103°C) have been reported (32,33). Similarly, HCl complexes undergo rapid formation and decomposition at –80°C (34).

Werner complexes can be used to form clathrates with the C₈ aromatic isomers (35–42). The aromatic compounds are released upon heating. Since the uptake and release characteristics of the four C₈ aromatic isomers are each different, this method has been suggested as a means of separating the isomers.

Inclusion compounds of the C₈ aromatic compounds with tris(*o*-phenylenedioxy)cyclotriphosphazene have been used to separate the individual isomers (43–47). The Schardinger dextrins, such as alpha-cyclodextrin, beta-dextrin, and gamma-dextrin are used for clathration; alpha-dextrin is particularly useful for recovering PX from a C₈ aromatic mixture (48,49). Pyromellitic dianhydride (50) and beryllium oxybenzoate (51) also form complexes, and procedures for separations were developed.

A 1:1 complex melting at 24.8°C is formed between PX and carbon tetrachloride (52). The other C₈ aromatic compounds do not form these complexes. Carbon tetrabromide and chloral (CCl₃CHO) form addition compounds with PX.

Manufacture of Xylenes

The initial manufacture of mixed xylenes and the subsequent production of high purity PX and OX consists of a series of stages in which (1) the mixed xylenes are initially produced; (2) PX and or OX are separated from the mixed xylenes stream; and (3) the PX- (and perhaps OX-) depleted xylene stream is isomerized back to an equilibrium mixture of xylenes and then recycled back to the separation step. These steps are discussed below.

Mixed Xylenes Production Via Reforming. Again, two principal methods for producing xylenes are catalytic reforming and toluene disproportionation. A general schematic for the production of PX and OX (along with benzene and toluene) via catalytic reforming is shown in Figure 3 (see BTX PROCESSING). In this example, a light fraction (ie, 65–175°C) from a straight run petroleum fraction or from an isocracker is fed to a catalytic reformer, unit A. This is followed by heart-cutting and extraction in units B, C, and D. The mixed xylenes stream must then be processed further to produce high purity PX and or OX. As discussed herein, high purity OX can be produced via distillation. However, because of the close boiling points of PX and MX, using distillation to produce high purity PX is impractical. Instead, other separation methods such as crystallization and adsorption are used. These separation processes are discussed herein.

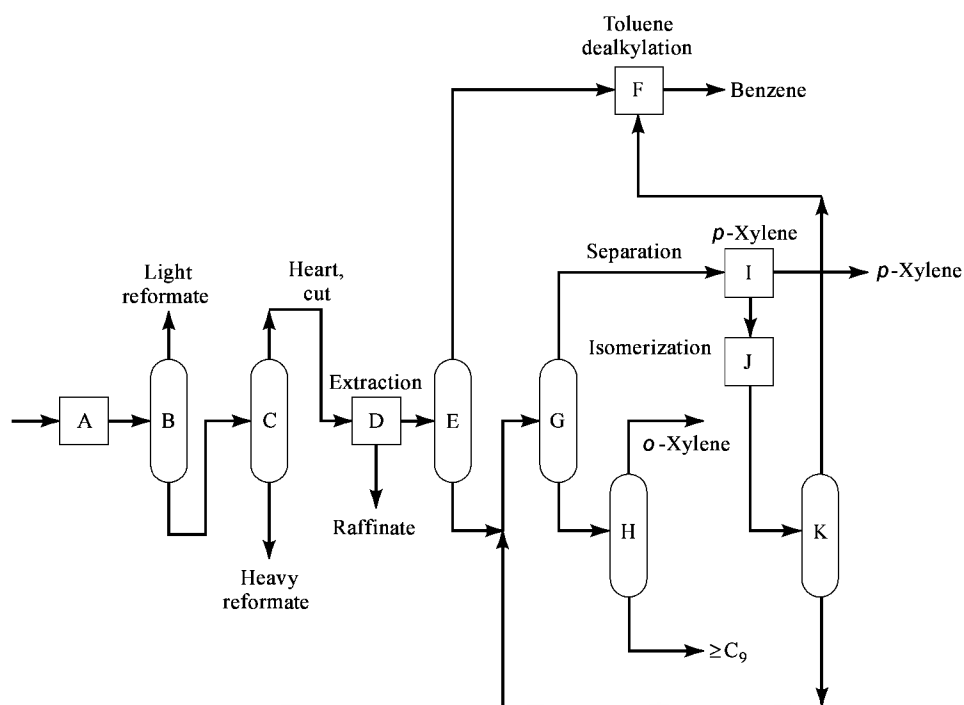
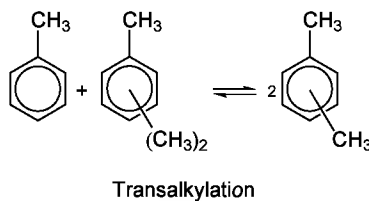
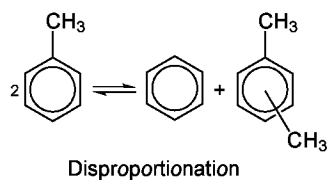


Fig. 3. General scheme for producing benzene, PX, and OX from catalytic reforming.

Xylenes Production Via Toluene Transalkylation and Disproportionation. The toluene that is produced from processes such as catalytic reforming can be converted into xylenes via transalkylation and disproportionation. Toluene disproportionation is defined as the reaction of 2 mol of toluene to produce 1 mol of xylene and 1 mol of benzene. Toluene transalkylation is defined as the reaction of toluene with C₉ or higher aromatics to produce xylenes:



Other species that are also present in the feed, such as ethylbenzene and methylethylbenzenes will also undergo transalkylation reactions. These reactions tend to approach an equilibrium that depends on the operating conditions.

There are several commercial processes that produce xylenes via disproportionation or transalkylation. These include UOP's Tatoray and PX-Plus, ARCO's Xylenes Plus, and Mobil's MTDP and STDP.

The Tatoray process was originally developed by Toray and is currently licensed by UOP (53–57). A schematic of the process is shown in Figure 4. In this process, toluene or a mixture of toluene and C_{10+} aromatics are reacted to form primarily xylenes and benzene. An equilibrium distribution of xylenes is produced. As shown in Table 4, the ratio of xylenes to benzene can be adjusted by altering the feed ratio of toluene to C_{10} aromatics.

Trimethylbenzenes are the preferred C₉ aromatic compound.

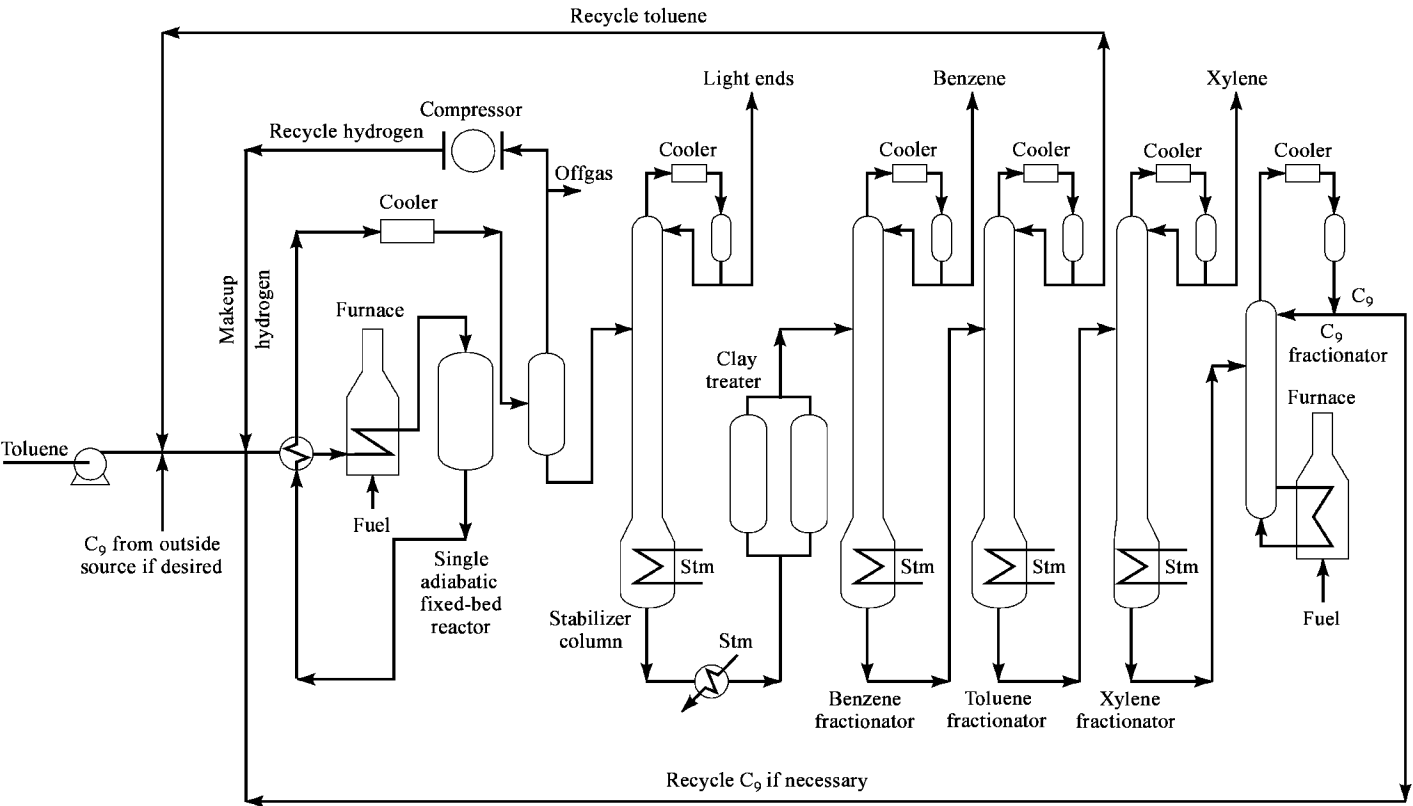


Fig. 4. UOP Tatoray process for xylenes production.

Table 4. Tatoray Transalkylation of Toluenes and C₉ Aromatic Compounds, Relative Wt Units

Feed		Product		
Toluene	C ₉ Aromatics	Xylenes	Benzene	Other
1000	0	460	425	125
500	500	675	200	125

Catalyst improvements have been made since the process was first developed. The current catalyst is designated TA-4 (57). It has a high per pass conversion and good stability. Yields to xylenes are reported to be over 97%. The pelleted catalyst is used in a fixed-bed reactor in the presence of hydrogen. Typical operating conditions are: 350–530°C, 1–5 MPa (10–50 atm), and H₂ hydrocarbon ratio of 5–12:1. Toluene per pass conversion is about 40–50%. Per pass conversion can be increased by increasing pressure or temperature, although this also reduces yield and increases the rate of catalyst deactivation. The product typically contains <2 mol% EB. The catalyst is regenerated by burning of the coke that is formed. The first commercial plant was built at Toray's Kawasaki plant in 1969. As of 1995, 32 units are reported to have been licensed (57).

The Xylenes Plus process also converts toluene with or without C₉ aromatics. This process was originally developed by ARCO and is also currently available for license from Lyondell and HRI IFP (58,59). Feeds that have been commercially used include catalytic reformat and hydrotreated pygas. The catalyst is reported to be a nonnoble metal catalyst. A moving bed system is used with a continuous catalyst regenerator. Hydrogen is not required for any facet (start-up, operation, or catalyst regeneration) of the process, and thus operating pressure is relatively low. The licensors report that operation at 30% toluene per pass conversion results in the highest possible yield of xylenes with minimal feedstock loss to gas or coke. EB production is close to zero. Liquid yields are reported to be 95–97 vol% with an equilibrium distribution of xylenes (PX OX MX 26 24 50). The yields of xylenes and benzene for various feed ratios of toluene and C₉ aromatics are given in Table 5. The first plant was installed at ARCO's (now Lyondell's) Houston refinery in 1968. To date, this process has been utilized at 5 plants worldwide.

Table 5. Xylenes Plus Transalkylation of Toluene and C₉ Aromatics Compounds, Relative Wt Units

Feed		Product		
Toluene	C ₉ Aromatic Compounds	Xylenes	Benzene	Other
1000	0	460	430	120
500	500	700	140	120

Mobil has developed several TDP processes. In the 1970s, Mobil developed their LTD (Low Temperature Disproportionation) process (60,61). This was a liquid-phase process which used what was described as siliceous zeolitic catalysts. Hydrogen was not required in the process. Reactor pressure was 4.5 MPa and WHSV of 0.68 kg oil / h kg catalyst. The initial reactor temperature was 127°C and was raised as the catalyst deactivated to maintain toluene conversion. The catalyst was regenerated after the temperature reached about 315°C. Regeneration consisted of conventional controlled burning of the coke deposit. The catalyst life was reported to be at least 1.5 yr.

In the mid-1970s, Mobil introduced their MTDP process (62). This is a vapor-phase process that uses ZSM-5-type zeolites. The process operates at 48 wt% toluene per pass conversion with a near equilibrium PX selectivity of 24% in the xylenes product. Typical operating conditions are inlet temperatures of 390–495°C, H₂ partial pressure of 4.1 MPa, H₂ hydrocarbon molar ratio of 4, and liquid hourly space velocity (LHSV) 1.0–2.0 h. EB production was about 3 wt% of feed. The first commercial use of this process was at Mobil's Naples, Italy petrochemical complex in 1975. In the 1980s, Mobil developed an improved ZSM-5 catalyst formulation for the MTDP process which was reported to have significant activity and stability advantages over the first catalyst. The higher activity of the improved catalyst meant reactor temperatures could be lowered by about 55°C, resulting in lower coking rates and longer catalyst cycle times. Hydrogen partial pressures could also be lowered. The latest version of MTDP is designated MTDP-3 (63). To date

six commercial units have been installed using MTDP technology.

In the late 1980s, Mobil commercialized the MSTDP (Mobil Selective Toluene Disproportionation) Process (64–66). In MSTDP, PX selectivity can be boosted to 80–90% at toluene conversion of about 25–30%. The significantly higher than equilibrium yield of PX is obtained by precoking the catalyst in the reactor. This reduces the pore size opening of the ZSM-5 catalyst and also eliminates acidity on the external surface which can cause reisomerization back to equilibrium. Inside the catalyst pores, an equilibrium mix of the three xylene isomers is obtained. However, the smaller molecular dimensions of PX compared to OX and MX translates into a relative diffusivity for PX that is several orders of magnitude higher than that of OX and MX. Because of its higher diffusivity, PX exits the catalyst pores more rapidly while the remaining OX and MX reequilibrate inside the pores to form more PX. The catalyst is initially selectivated in the reactor by precoking with toluene. This pretreatment is performed at the beginning of each cycle. As coking occurs, the toluene conversion progressively drops from 60% to about 20–30%, whereas the PX selectivity increases from 24% to about 80–90%. The higher PX concentration produced in the MSTDP process compared to conventional TDP processes means that the subsequent crystallization or adsorption PX recovery facilities can be significantly smaller. Typical operating conditions are: 400–470°C, 2.0–3.4 MPa, WHSV 2–4 h, H₂ hydrocarbon molar ratio of 1–3:1. The MSTDP process was first demonstrated at Enichem's refinery in Gela, Italy in 1988. As of mid-1996, 9 units had been licensed. Of these, six are currently operating, including a reformer retrofit by Exxon and a grass roots unit by Koch in the United States, while three others were scheduled for start-up before the end of 1997.

In 1995, Mobil announced an improved MSTDP process named MTPX (Mobil Toluene to Paraxylene) (67). Although there are no technical details of this process in the technical literature at the present time, at least some of the improvement may consist of doing the catalyst pore selectivation with the organosilicones that are the subject of a recent Mobil Patent (68). Unlike the coking selectivation, this treatment could be done during catalyst manufacture. Mobil has also announced that it does not plan to license this technology and no longer plans to offer new MSTDP licenses (69). Mobil has built a grassroots MTPX plant at its Beaumont plant.

In 1997, UOP announced the PX-Plus process which also uses a selectivated catalyst to convert toluene to para-rich xylenes. Fina commercialized a TDP process known as the (T2PX) process in 1984 (70). It uses a proprietary catalyst to react toluene at 42–48% conversion with selectivities to benzene of 42 wt % and to xylenes of 46 wt %. The xylenes produced are at equilibrium. Typical commercial operating conditions of 390–495°C, H₂ partial pressure of 4.1 Mpa, H₂ hydrocarbon molar ratio of 4:1, and LHSV of 1–2 h. Fina's first commercial implementation occurred in 1985 at their Port Arthur refinery.

Separation Processes for PX. There are essentially two methods that are currently used commercially to separate and produce high purity PX: (1) crystallization and (2) adsorption. A third method, a hybrid crystallization–adsorption process, has been successfully field-demonstrated and the first commercial unit is expected in the near future.

Crystallization. Low temperature fractional crystallization was the first and for many years the only commercial technique for separating PX from mixed xylenes. As shown in Table 2, PX has a much higher freezing point than the other xylene isomers. Thus, upon cooling, a pure solid phase of PX crystallizes first. Eventually, upon further cooling, a temperature is reached where solid crystals of another isomer also form. This is called the eutectic point. PX crystals usually form at about 4°C and the PX-MX eutectic is reached at about 68°C. In commercial practice, PX crystallization is carried out at a temperature just above the eutectic point. At all temperatures above the eutectic point, PX is still soluble in the remaining C₈ aromatics liquid solution, called mother liquor. This limits the efficiency of crystallization processes to a per pass PX recovery of about 60–65%.

The solid PX crystals are typically separated from the mother liquor by filtration or centrifugation. Good solid–liquid separation is important for obtaining high purity PX. One key to good separation is crystal size. The larger the crystal, the better the separation. Crystal size is affected by the degree of supersaturation and nucleation, which in turn is affected by a number of parameters, including temperature, agitation, and the presence of crystal growth sites.

PX crystals are typically produced in two or more stages of crystallization, separated by centrifuges. Commercial crystallizers use either direct contact or indirect refrigeration. The latter has the disadvantage that the walls of the cooled surface tend to foul, which reduces heat transfer. The first crystallizer stage is usually at the lowest temperature. The cake from this stage has a purity of about 80–90%. The impurity arises from the mother liquor which wets the crystal surface or is occluded in the crystal cake. The efficiency of the solid–liquid separation depends on the temperature and the loading of the centrifuges. As temperature falls, the viscosity and density of the mother liquor rise sharply. Thus, it becomes more difficult for the centrifuges to achieve effective separation.

In the second crystallizer stage, the crystals are usually reslurried with a higher purity PX stream from a later stage of purification. A second stage of centrifugation is sufficient in most cases to give PX purity >99%.

Currently, about 40% of the PX produced worldwide uses crystallization technology. A number of crystallization processes have been commercialized over the years. The more common ones are those developed by Chevron, Krupp, Amoco, ARCO (Lyondell), and Phillips. Some of the features of these processes are discussed herein.

The Chevron process (71) is shown in Figure 5. It consists of two crystallizers in series operated at difficult pressures. Direct contact cooling is used. This is accomplished by injecting liquid CO₂ with the feed to the crystallizer. As the slurry rises, part of the CO₂ vaporizes, causing the temperature to drop below the saturation temperature, and crystallization occurs. Because cooling is gradual, the degree of supersaturation is low and thus crystal growth occurs on the existing crystals. This leads advantageously to the formation of relatively large crystals, rather than many small ones. The crystals and slurry move down from the crystallizer body. Most of the slurry is recycled, but some is withdrawn and sent to the second crystallizer, which is operated under vacuum. The operation of the second crystallizer is similar to the first, except that typically it is not necessary to inject additional CO₂. The crystals are separated from the mother liquor in two stages. The first stage uses screen bowl centrifuges, and the second uses pusher centrifuges. The Chevron process offers the advantage that large crystals are obtained in a relatively short residence time, which permits good solid–liquid separation in the centrifuges. Plants using this process have been licensed and built in the United States, Germany, Mexico, Japan, and in the United Kingdom.

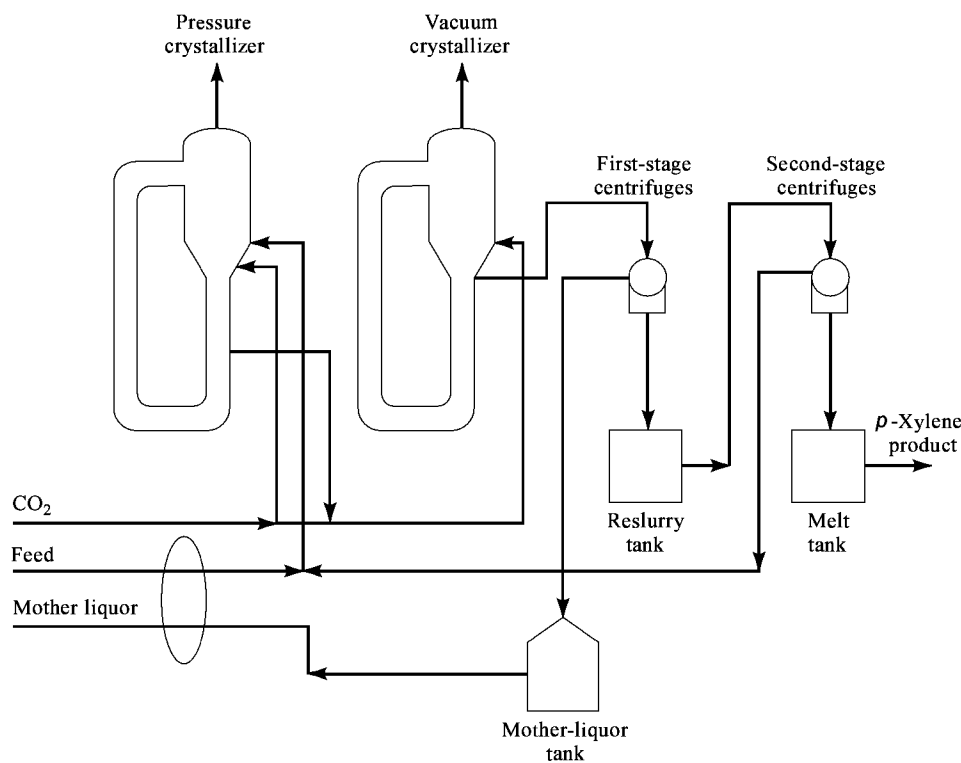


Fig. 5. Chevron PX crystallization process.

The Krupp process for PX crystallization uses scraped chillers for crystallization (72). Rotary drum filters and centrifuges are used for phase separation. The first-stage solids are reslurred in a PX-rich filtrate from the final PX product centrifuge. Another feature is that the second-stage mother liquor is further chilled and centrifuged before being passed back to the first-stage feed. The resultant solids are added to the first-stage crystals.

The Amoco PX crystallization process (73,74) is a two-stage process that operates with indirect cooling. A schematic of this process is shown in Figure 6. Ethylene is used as the coolant in the first stage and propane is used in the second stage. In the first-stage crystallizer, the temperature is brought down in stages to near the PX–MX eutectic. The first stage cake is melted and sent to a second-stage crystallizer, which is designed like the first, but uses propane refrigerant instead of ethylene. The crystallizers are fitted with scrapers mounted on a central shaft, which provides agitation and maintains a good heat-exchange surface. The residence time in each of the two crystallizers is about 3 h, in order to encourage crystal growth.

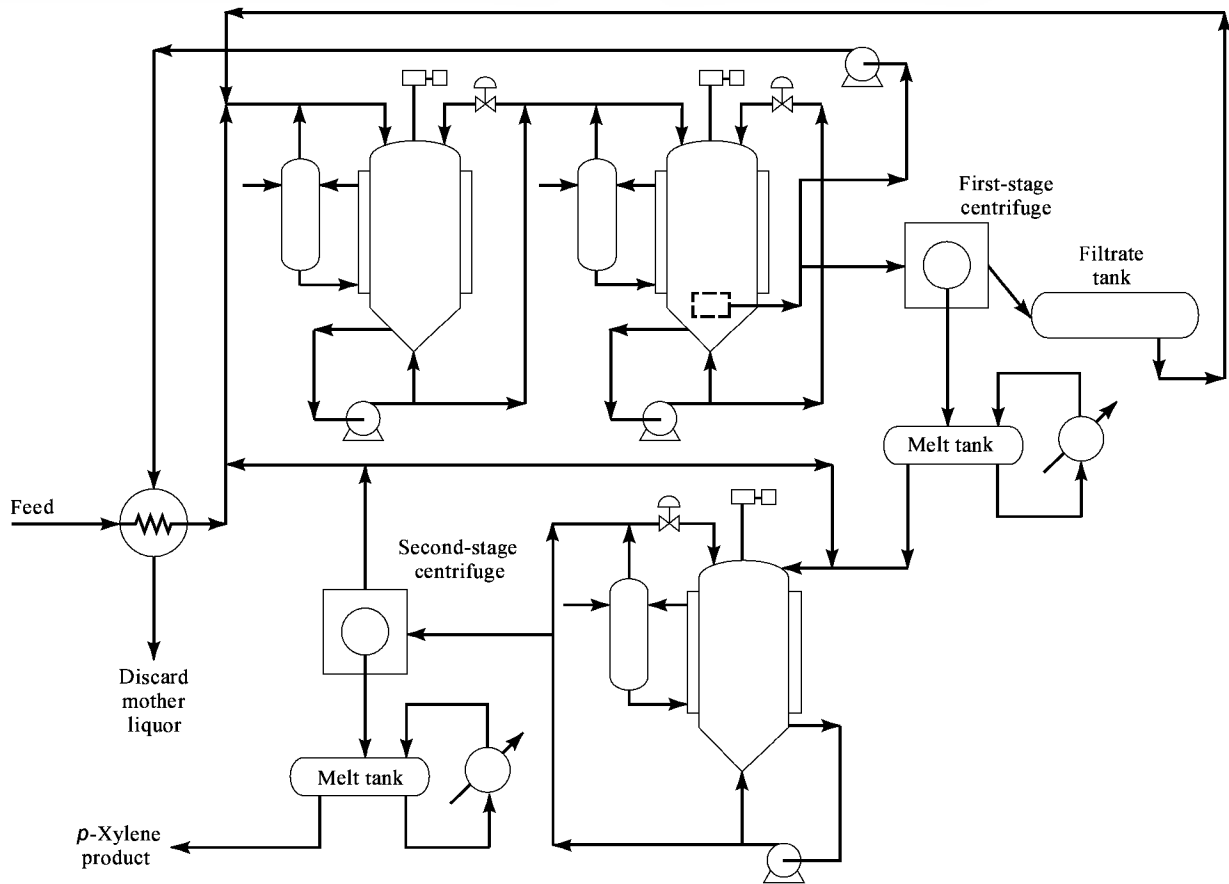


Fig. 6. Amoco PX crystallization process.

The ARCO process is similar to the Amoco process in that it is a two-stage crystallization process that uses external ethylene refrigerant in the first stage and propane in the second stage (75,76). The first stage consists of two crystallizers which operate at 40 to 50°C and 55 to 70°C, respectively. The first-stage slurry flows to a continuous centrifuge where crystals of 85–90% purity are removed from the filtrate. The crystals are remelted and then sent to a single second-stage crystallizer operated at 20 to 0°C, depending on the feedstock and desired purity. The crystal slurry flows from the crystallizer to a second-stage continuous centrifuge. The crystals are washed with toluene in the centrifuge and melted. The liquid is then fed to a

PX-toluene splitter where high purity PX is taken as a bottoms product. Eight plants have been built worldwide.

The Maruzen process uses ethylene gas as a direct refrigerant in a two-stage process (77,78). The first stage slurry is centrifuged, partially remelted, and fed to the second centrifuge.

The Phillips process is a two-stage crystallization process that uses a pulsed column in the second stage to purify the crystals (79,80). In the pulsed column, countercurrent contact of the high purity PX liquid with cold crystals results in displacement of impurities. In the first stage, a rotary filter is used. In both stages, scraped surface chillers are used. This process was commercialized in 1957, but no plants in operation as of 1996 use this technology.

Relatively recently, several companies have commercialized static crystallization processes based on progressive freezing. In progressive freezing, crystals are allowed to form on a cooled surface until a certain proportion of the original batch is frozen. The remaining melt is drained away from the solid layer. In some versions, the frozen layer is also gently warmed to sweat out impurities caught in the crystal structure. Finally, the PX is recovered by heating the surface to completely melt off the frozen layer. In systems developed by BEFS and Sulzer Chemtech, the crystal layer is grown on vertical cooled plates. In the BEFS PROKEM PROABD process, feed containing about 80–99.5% PX is further purified to 99.9 ± % in a single-stage melt static crystallizer (81). After crystallization, the crystalline mass is purified by a partial melting of the crystals to wash out adhering impurities. This process can be used as a finishing step to further improve the purity of PX product from adsorption, crystallization, or MSTDP units. An 8000-t/yr plant has been in operation for three years in Gonfreville, France. BEFS Prokem has recently won two contracts to implement the process at Reliance Industries in India and Neftochim E.A.D. in Bulgaria (82).

One drawback of static crystallization is that crystal layer growth rates are very slow. In the Sulzer MWB process, growth rates are greatly improved by allowing a film to flow down vertical tubes (83).

Adsorption Processes. Adsorption represents the second and newer method for separating and producing high purity PX. In this process, adsorbents such as molecular sieves are used to produce high purity PX by preferentially removing PX from mixed xylene streams. Separation is accomplished by exploiting the differences in affinity of the adsorbent for PX, relative to the other C_8 isomers. The adsorbed PX is subsequently removed from the adsorbent by displacement with a desorbent. Typical PX recovery per pass is over 95%, compared to only 60–65% for crystallization. Thus recycle rates to the separation and isomerization units are much smaller where adsorption is used.

Currently, there are three commercially available PX adsorption processes: UOP's Parex, IFP's Eluxyl, and Toray's Aromax (not to be confused with Chevron's Aromax process for reforming naphtha into aromatics). In all of these processes, the feed and desorbent inlets and the product outlet ports are moved around the bed, simulating a moving bed.

In 1971, UOP first commercialized their PX adsorbent process, Parex (84–94). It is one of several adsorption processes that constitutes UOP's SORBEX adsorption technologies. Ideally, an adsorption process would be like the hypothetical moving-bed adsorption system in Figure 7a. In this system, solid adsorbent moves through the bed countercurrently to the liquid. The feed is assumed to be a binary mixture of component A (PX) and component B (MX, OX, and EB). Component A is selectively adsorbed. Desorbent D enters the top of the bed. Extract (A + desorbent) is withdrawn at a point between the desorbent and feed injection points. Raffinate containing B + desorbent is withdrawn below the feed injection point. The adsorber is effectively divided into four separate zones. In zone 1, feed component A is completely adsorbed from the downward-moving liquid stream. Zone 2 is a rectification section where component B is completely desorbed from the adsorbent. Zone 3 is a desorbent section where A is displaced from the adsorbent by the desorbent D. Zone 4 is a secondary rectification section where component B is readsorbed, preventing B from entering zone 3 where it would contaminate the product A. The liquid composition in these four zones is illustrated in Figure 7b.

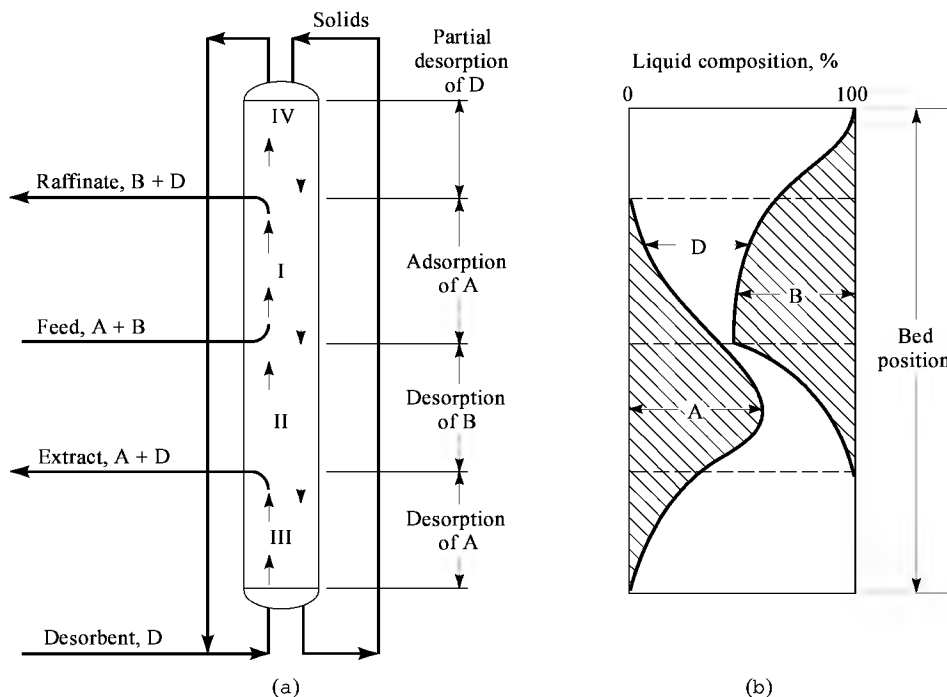


Fig. 7. Ideal PX adsorption process.

Courtesy of UOP.

There are, however, several difficulties which prevent this ideal system from being literally implemented commercially. These include likely mechanical erosion of the adsorbent if it were to circulate in the bed and the difficulty of obtaining truly uniform flow of both solids and liquids in large-diameter beds. However, UOP recognized that the same results can be accomplished on a commercial scale by holding the bed stationary, but periodically moving the positions at which the various liquid streams enter and leave. By shifting the positions of feed-injection and product-withdrawal streams in the direction of flow through the bed, the countercurrent movement of solid is simulated. Although from a practical standpoint the fluid-injection and withdrawal positions cannot be moved continuously, nearly the same effect can be achieved by providing multiple liquid access lines to the bed and periodically switching each stream to the adjacent port, always maintaining the same relative distance between the various streams.

The flow diagram illustrating the commercial implementation of this fixed-bed continuous adsorption process by UOP is shown in Figure 8. The feed, desorbent, and product ports are continuously changed, using a patented rotary valve. In the particular example shown in Figure 8, 12 lines are connected to the rotary valve. Only four lines are active at any one instant. Desorbent is injected into one line, feed is injected into another, extract is withdrawn from another, and raffinate is withdrawn from still another. By rotating the valve by one position, a different set of four ports is activated. Thus a condition is simulated whereby the solid appears to be moving past fixed portions of fluid feed, product, and withdrawal. The operating conditions are 250–400°C and moderate pressures. Earlier versions of Parex used a light desorbent, toluene. More recent versions have used a heavier desorbent, usually *p*-diethylbenzene, which has resulted in lower energy costs associated with separating the PX from the desorbent. The newest adsorbent, ADS-27 was

commercialized in 1990. Compared to its predecessor, ADS-7, UOP claims that capacity has increased 15%. Another reported benefit is that ADS-27 is manufactured and loaded into the commercial unit at the moisture level required for operation. This eliminates the need for the adsorbent to go through the initial vapor-phase benzene dry-down step required by previous adsorbents. The first Parex unit was commercialized in 1971. As of 1995, 59 units had been licensed (94).

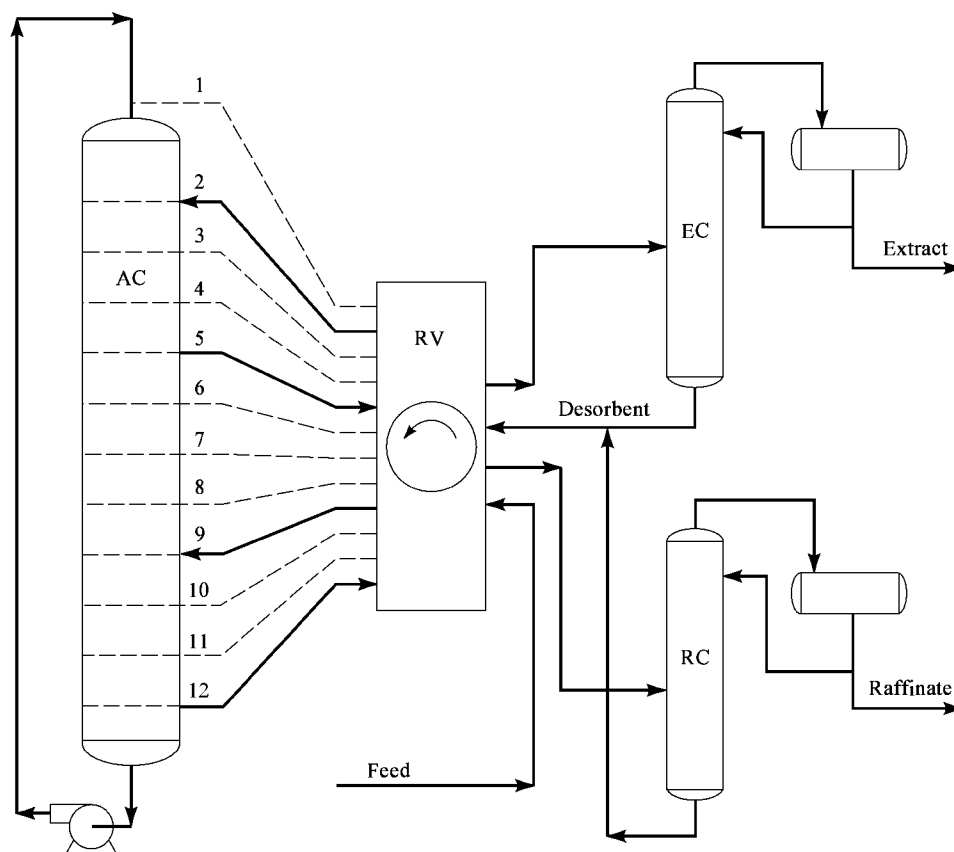


Fig. 8. UOP Parex simulated moving bed for adsorptive separation. AC = adsorbent chamber; RV = rotary valve; EC = extract column; RC = raffinate column. Lines: 2-desorbent; 5-extract; 9-feed; 12-raffinate. All other ports are closed at this time.

The Aromax process was developed in the early 1970s by Toray Industries, Inc. in Japan (95–98). The adsorption column consists of a horizontal series of independent chambers containing fixed beds of adsorbent. Instead of a rotary valve, a sequence of specially designed on–off valves under computer control is used to move inlet and withdrawal ports around the bed. Adsorption is carried out in the liquid phase at 140°C, 785–980 kPa, and 5–13 L/h. PX yields per pass is reported to exceed 90% with a typical purity of 99.5%. The first Aromax unit was installed at Toray's Kawasaki plant in March 1973. In 1994, IFP introduced the Eluxyl adsorption process (59,99). The proprietary adsorbent used is designated SPX 3000. Individual on–off valves controlled by a microprocessor are used. Raman spectroscopy is used to measure concentration profiles in the column. A 10,000 t/yr demonstration plant was started and successfully operated at Chevron's Pascagoula plant from 1995–96. IFP has licensed two hybrid units.

Asahi Chemical Industry Company Ltd. was working to develop an adsorption process in the late 1970s and early 1980s that was to produce high purity EB as well as PX (100–103). In 1981 they reported that pilot plant results were being confirmed in larger equipment. However, this process does not appear to have been commercialized.

Hybrid Crystallization/Adsorption Process. In 1994, IFP and Chevron announced the development of a hybrid process that reportedly combines the best features of adsorption and crystallization (59,99). In this option of the Eluxyl process, the adsorbent bed is used to initially produce PX of 90–95% purity. The PX product from the adsorption section is then further purified in a small single-stage crystallizer and the filtrate is recycled back to the adsorption section. It is reported that ultrahigh (99.9+%) purity PX can be produced easily and economically with this scheme for both retrofits of existing crystallization units as well as grass-roots units. A demonstration plant was built at Chevron's Pascagoula refinery in 1994.

MX Separation Process. The Mitsubishi Gas–Chemical Company (MGCC) has commercialized a process for separating and producing high purity MX (104–113). In addition to producing MX, this process greatly simplifies the separation of the remaining C₈ aromatic isomers. This process is based on the formation of a complex between MX and HF–BF₃. MX is the most basic xylene and its complex with HF–BF₃ is the most stable. The relative basicities of MX, OX, PX, and EB are 100, 2, 1, and 0.14, respectively.

MX of >99% purity can be obtained with the MGCC process with <1% MX left in the raffinate by phase separation of hydrocarbon layer from the complex–HF layer. The latter undergoes thermal decomposition, which liberates the components of the complex.

A schematic of the MGCC process is shown in Figure 9. The mixed C₈ aromatic feed is sent to an extractor (unit A) where it is in contact with HF–BF₃ and hexane. The MX–HF–BF₃ complex is sent to the decomposer (unit B) or the isomerization section (unit D). In the decomposer, BF₃ is stripped and taken overhead from a condenser–separator (unit C), whereas HF in hexane is recycled from the bottom of C. Recovered MX is sent to column E for further purification. The remaining C₈ aromatic compounds and hexane are sent to raffinate column F where residual BF₃ and HF are separated, as well as hexane for recycle. Higher boiling materials are rejected in column H, and EB and OX are recovered in columns I and J. The overhead from J is fed to unit K for PX separation. The raffinate or mother liquor is then recycled for isomerization.

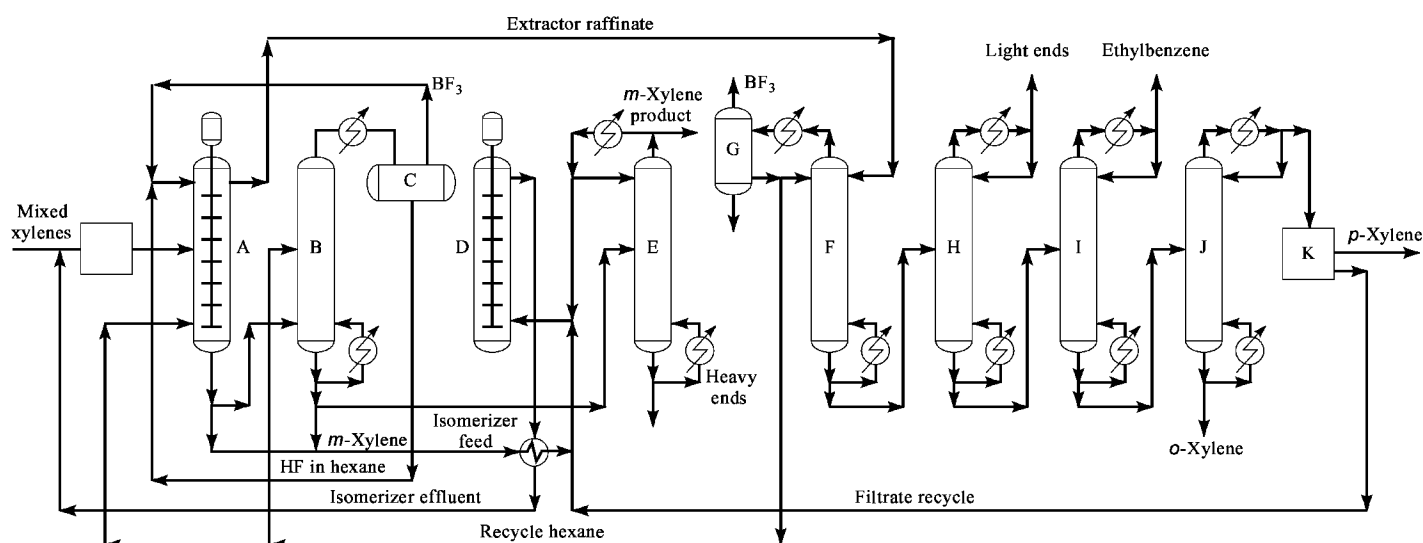


Fig. 9. Xylenes separation via Mitsubishi Gas-Chemical Co. HF-BF₃ extraction-isomerization process (107). A = extractor; B = decomposer; C = separator; D = isomerization reactor; E = heavy ends tower; F = raffinate tower; G = separator; H = light ends fractionator; I = ethylbenzene fractionator; J = OX fractionator; K = PX crystallizer.

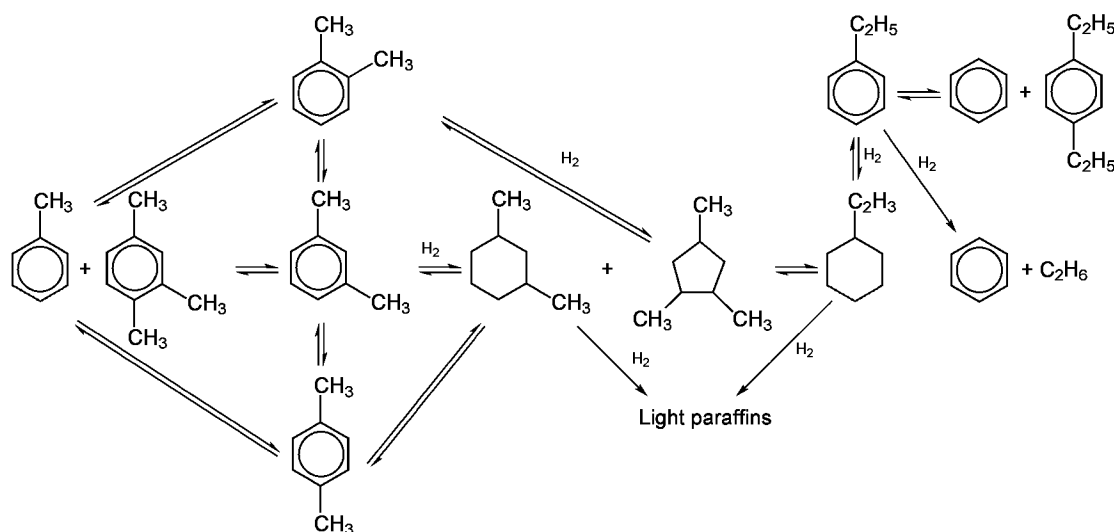
The MGCC process is used in Japan, the United States, and Spain.

Xylene Isomerization. After separation of the preferred xylenes, ie, PX or OX, using the adsorption or crystallization processes discussed herein, the remaining raffinate stream, which tends to be rich in MX, is typically fed to a xylenes isomerization unit in order to further produce the preferred xylenes. Isomerization units are fixed-bed catalytic processes that are used to produce a close-to-equilibrium mixture of the xylenes. To prevent the buildup of EB in the recycle loop, the catalysts are also designed to convert EB to either xylenes, benzene and lights, or benzene and diethylbenzene.

Historically, the isomerization catalysts have included amorphous silica-aluminas, zeolites, and metal-loaded oxides. All of the catalysts contain acidity, which isomerizes the xylenes and if strong enough can also crack the EB and xylenes to benzene and toluene. Dual functional catalysts additionally contain a metal that is capable of converting EB to xylenes.

The three major commercial licensors of xylenes isomerization processes are Engelhard, UOP, and Mobil. Several other companies have developed and used their own catalysts. These companies include Mitsubishi Gas-Chemical, Toray, ICI, Amoco, and Shell. All of these processes are discussed herein.

Dual Function Catalytic Processes. Dual-function catalytic processes use an acidic oxide support, such as alumina, loaded with a metal such as Pt to isomerize the xylenes as well as convert EB to xylenes. These catalysts promote carbonium ion-type reactions as well as hydrogenation-dehydrogenation. In the mechanism for the conversion of EB to xylenes shown, EB is converted to xylenes



by hydrogenation of the aromatic ring and formation of naphthenic intermediates such as ethylcyclohexane, dimethylcyclohexanes, and trimethylcyclohexanes. To provide an effective pathway between EB and xylenes, a certain concentration of the naphthenic intermediates must be maintained. This is accomplished by proper selection of operating conditions and separation and recycle of the naphthenic intermediates. Side reactions include disproportionation and hydrodealkylation of aromatic compounds, and hydrodealkylation of naphthenes. Effective catalysts minimize the loss of xylenes and EB via these side reactions.

Commercial processes which use a dual-functional catalyst are Octafining, Isomar, and Isolene.

The Octafining process (114–116) was developed and commercialized by Atlantic Richfield and Engelhard in the early 1960s. The first-generation catalyst was prepared by mixing equal amounts of a silica-alumina cracking catalyst with Pt on alumina. The Pt content of the mixture was about 0.5 wt %. The EB approach to equilibrium was 88%, with an 80% selectivity to xylenes. Reaction conditions consist of temperature of 425–480°C; pressure of 1.14–2.51 MPa; H₂ hydrocarbon ratio < 10 : 1 (preferably 4–6:1); and LHSV = 0.6–1.6 h. An equilibrium mixture of xylenes is produced. To maintain conversion, the reaction temperature is gradually increased as the catalyst deactivates up to a maximum temperature of about 480°C. The catalyst is then regenerated. The first Octafining unit was placed onstream in 1960. As of 1995, 30 units had been licensed in 17 countries (99).

Recently, a new second-generation catalyst has been commercialized. This new catalyst, O-750, is formulated by loading Pt on alumina and combining the mixture with H-mordenite such that the final Pt content is <0.5 wt% (101). This catalyst, along with several other process improvements, forms the basis of the Octafining II process. Compared to Octafining I, yields and per pass conversions have been improved. Catalyst cycle lengths typically exceed 3 yr. The catalyst is capable of cracking C₉ nonaromatics, thereby increasing xylene purity. IFP has recently acquired the rights to license the Octafining II process worldwide (99).

UOP's Isomar process (56,117–119) was originally developed to use dual-functional catalysts. The first-generation catalyst contained Pt and halogen on alumina. Operating conditions using this catalyst were 399°C; 1.25 MPa; 2 LHSV; and H₂ hydrocarbon ratio of 6:1. A C₈ naphthene concentration of

2–9° o was maintained in the process loop. In the mid-1980s UOP introduced an improved catalyst, I-9. Unlike the original catalyst, this catalyst does not require chlorine addition. Process conditions were modified to 388°C and 1.68 MPa.

In 1993, UOP commercialized an improved Pt-based catalyst, I-210. This catalyst is based on a molecular sieve, but not an aluminosilicate zeolite. UOP claims that yields are about 10° o better than those for I-9 catalyst. EB to xylenes conversion is about 22–25° o with a C₈ aromatics per pass loss of about 1.2–1.5° o. As discussed below, UOP's Isomar process can also use zeolite catalysts which convert EB to benzene rather than to xylenes. UOP has licensed over 40 Isomar units.

The Isolene II process was commercialized in 1971 by Toray Industries (120–122). The catalyst is Pt on an acidic support. Operating conditions are reported to be 250–500°C and 1–3 MPa. The first Isolene II plant was built at Toray's Kawasaki complex.

Zeolite and Molecular Sieve-Based Process. Mobil has commercialized several xylene isomerization processes that are based on ZSM-5. Amoco has developed a process based on a medium-pore borosilicate molecular sieve.

Mobil's Low Pressure Isomerization Process (MLPI) was developed in the late 1970s (123,124). Two unique features of this process are that it is operated at low pressures and no hydrogen is used. In this process, EB is converted to benzene and diethylbenzene via disproportionation. The patent believed to be the basis for the MLPI process (123) discusses the use of H-ZSM-5 zeolite with an alumina binder. The reaction conditions described are start-of-run temperatures of 290–380°C, a pressure of 273 kPa; and WHSV of 5–8.5 h. The EB conversion is about 25–40° o depending on reaction conditions, with xylene losses of 2.5–4° o. The PX approach to equilibrium is about 99–101° o. The first commercial unit was licensed in 1978. A total of four commercial plants have been built.

A second Mobil process is the Mobil's Vapor Phase Isomerization Process (MVPI) (125,126). This process was introduced in 1973. Based on information in the patent literature (125), the catalyst used in this process is believed to be composed of NiHZSM-5 with an alumina binder. The primary mechanism of EB conversion is the disproportionation of two molecules of EB to one molecule of benzene and one molecule of diethylbenzene. EB conversion is about 25–40° o, with xylene losses of 2.5–4° o. PX is produced at concentration levels of 102–104° o of equilibrium. Temperatures are in the range of 315–370°C, pressure is generally 1480 kPa, the H₂ hydrocarbon molar ratio is about 6:1, and WHSV is dependent on temperature, but is in the range of 2–50, although normally it is 5–10.

Mobil's High Temperature Isomerization (MHTI) process, which was introduced in 1981, uses Pt on an acidic ZSM-5 zeolite catalyst to isomerize the xylenes and hydrodealkylate EB to benzene and ethane (126). This process is particularly suited for unextracted feeds containing C₈, aliphatics, because this catalyst is capable of cracking them to light paraffins. Reaction occurs in the vapor phase to produce a PX concentration slightly higher than equilibrium, ie, 102–104° o of equilibrium. EB conversion is about 40–65° o, with xylene losses of about 2° o. Reaction conditions are temperature of 427–460°C, pressure of 1480–1825 kPa, WHSV of 10–12, and a H₂ hydrocarbon molar ratio of 1.5–2:1. Compared to the MVPI process, the MHTI process has lower xylene losses and lower formation of heavy aromatics.

In the early 1990s, Mobil commercialized their High Activity Isomerization (MHAI) process (124,127,128). A patent corresponding to this process (127) describes the use of a two catalyst system in which the catalyst beds are separated from each other but can be located in the same reactor. The first catalyst consists of a noble metal (preferably Pt) on ZSM-5 zeolite having a crystal size of at least 1 micron with an alpha value >100. The primary function of this catalyst is to hydrodealkylate the EB to benzene and ethane. The second catalyst consists of a noble metal, preferably Pt, on a ZSM-5 zeolite having a crystal size <1 micron and an alpha value <100. The primary function of this catalyst is to isomerize the xylenes. Preferred operating conditions are reported to be 400–480°C; 450–2860 kPa, WHSV of 3–50; and a H₂ hydrocarbon molar ratio of 1–5:1. For a nonextracted feed containing 15° o EB, typical EB conversions are 65° o, with xylene losses of 1.8° o. Compared to MHTI, for a given EB conversion MHAI xylene losses are lower and nonaromatics conversion higher. As of 1995, 9 MHAI units had been commercialized (127).

Amoco has developed a series of xylene isomerization catalysts called AM-SAC catalysts (129). These catalysts are based on a medium-pore borosilicate molecular sieve, designated AMS-1B. They are reported to retard xylenes destruction by suppressing the transfer of methyl groups. Amoco claims that the required intermediate is too bulky to form within the pores of AMS-1B. EB is reacted via transfer of ethyl groups in a reaction scheme that involves dealkylation, formation of ethylene, and realkylation. Optionally, a hydrogenating metal such as Pt is added to inhibit realkylation by intercepting the ethylene and converting it to ethane, which is essentially unreactive. This has the advantage of further limiting xylene losses by preventing the alkylation of xylenes with C₂ groups to ethyl methyl benzenes. Amoco claims to have used its AMSAC catalysts in its PX units since 1980.

Amorphous Silica–Alumina Based Processes. Amorphous silica–alumina catalysts had been used for many years for xylene isomerization. Examples are the Chevron (130), Maruzen (131), and ICI (132–135). The primary advantage of these processes was their simplicity. No hydrogen was required and the only side reaction of significance was disproportionation. However, in the absence of H₂, catalyst deactivation via coking was rapid and regeneration was required every 3–30 days. Swing reactors were used to permit operation to continue during catalyst regenerations. The volume of catalyst used was fairly large, although the catalyst was inexpensive. Losses via disproportionation tended to be high, up to 5–10° o xylene and 10–15° o EB per pass under conditions at which xylenes were isomerized to near-equilibrium values. PX was produced at concentrations of 95–100° o of equilibrium.

Freshly regenerated catalyst was brought onstream at 370–420°C and temperature was gradually increased to maintain a >95% approach to equilibrium. The cycle length depended on feed and was longer when the OX concentration in the feed was low. The catalysts were very stable to regeneration by burnoff of coke and typically lasted several years before needing to be replaced. Generally, the catalyst activity loss was the result of a decrease in surface area which occurred slowly at regeneration conditions.

The ICI process was used in four U.K. plants and plants in Poland, Japan, and the former Soviet Union. This process operated without any feed diluent and the catalyst could undergo at least one rejuvenation *ex situ*. The temperature operating range was 430–450°C, with carbon deposition increasing rapidly at higher temperatures. Cycle lengths tended to be three to four days before regeneration. In most instances, EB was rejected from the fresh feed although concentrations of 20–25° o could be tolerated. The feed rate was about 0.5 h LHSV over that of a compressed pellet catalyst.

The Chevron process was used in two U.S. plants, although it is no longer used. Cycle lengths ranged from 6–30 d, depending on catalyst age and OX content of the feed. Operating conditions were temperature of 370–470°C and space velocity of about 0.5 h. Addition of 5 wt ° o steam reduced disproportionation losses.

Economic Aspects

Worldwide capacities as of 1992 for the production of the C₈ aromatic isomers are summarized in Table 6. U.S. capacity and production figures are given in Table 7.

Table 6. Worldwide Capacity for C₈ Aromatic Isomers in 1995,^a 10³ t

Location	OX	PX	MX	EB
United States	508	3,244	110	5,944
Canada and Mexico	122	608	0	1,104
Western Europe	665	1,310	0	5,157
Eastern Europe	674	852	0	1,841
Japan	302	2,200	55	3,587
other Asia	503	3,306	0	2,644
other	282	438	0	1,158

<i>Total</i>	<i>3,056</i>	<i>11,958</i>	<i>165</i>	<i>21,435</i>
--------------	--------------	---------------	------------	---------------

^a Ref. 136.

Table 7. U.S. Xylenes Economic Data in 1995,^a 10³ t

Parameter	Mixed xylenes	OX	MX	PX
production	4290	460	98	2877
imports	448	41	5	296
exports	616	78	32	615
consumption	4122	423	71	2558
OX	460			
MX	98			
PX	3452			
phthalic anhydride		363		
isophthalic acid			61	
dimethyl terephthalate				745
terephthalic acid				1743
other	112	60	8	70
year-end capacity	6140	508	110	3244
production capacity	0.70	0.91	0.89	0.89

^a Ref. 137.

U.S. producers and capacities for PX, OX, and EB are summarized in Table 8. Amoco is the only producer of MX with an annual capacity of 110,000 t. In recent years, many small xylene plants have shut down because the operations were not economic or the technology being used was obsolete.

Table 8. U.S. Capacity for PX, OX, and EB, 1995,^a 10³ t

Company	Location	Xylene process		PX	OX	EB
		Separation ^b	Isomerization			
Amoco	Decatur, Ala.	Amoco	Amoco	651		
	Texas City, Tex.	Amoco	Amoco	685		412
Arco	Channelview, Tex.					1,265
Chevron	Pascagoula, Miss.	Chevron	Mobil	242		
	St. James, La.					771
CosMar	Carville, La.					998
Dow	Freeport, Tex.					794
Exxon	Baytown, Tex.	Parex ^c	Mobil	454	127	
Huntsman	Bayport, Tex.					562
Koch	Corpus Christi, Tex.	Parex ^c	Mobil	590	125	
Lyondell	Houston, Tex.	Arco	Mobil	181	109	
Mobil	Chalmette, La.	Amoco	Mobil	70	70	
Phillips	Guayama, P.R.	Parex ^c	Mobil	371	77	
Rexene	Odessa, Tex.					159
Sterling	Texas City, Tex.					816
Westlake	Lake Charles, La.					167
<i>Total</i>				<i>3,244</i>	<i>508</i>	<i>5,944</i>

^a Ref. 136.

^b Crystallization process, unless otherwise noted.

^c Adsorption process.

The price histories for the four isomers are given in Reference 1. The average values started to increase drastically in late 1973 owing to increases in world crude oil prices. They also increased sharply from 1979–1985 in response to increased crude oil prices and shortages of xylenes. From 1986–1992 the prices generally declined as availability increased and crude oil prices declined, except in 1989 when derivatives demand was strong and xylenes supply was tight.

Analytical Methods

Typical ASTM methods for analyzing the C₈ aromatic isomers are listed in Table 9.

Table 9. Analytical Methods for C₈ Aromatic Isomers

Assay	ASTM method
purity, mol %	D1016
distillation range, °C	D850
specific gravity, 15.5°C 15.5°C	D891
color Saybolt	D156
flash point, °C	D56

p-Xylene purities can be determined by both freezing point and chromatographic methods. One advantage of freezing point methods is that they can detect the presence of water and carbon dioxide impurities, which chromatographic methods do not. Chromatographic methods exhibit another disadvantage, ie, for the analysis of very high purity PX streams, in that the PX peak tends to tail over the MX peak. Conditions for one chromatographic

method for separating C₈ aromatic isomers are given in Table 10. This procedure readily separates benzene, toluene, and the C₈ aromatic isomers. Di-*n*-propyl tetrachlorophthalate is also a useful column coating because it reverses the positions of PX and MX. Thus, in the case of this column, MX elutes ahead of PX and provides better analysis for streams containing high purity PX.

Table 10. Chromatographic Separation of C₈ Aromatic Isomers

Parameter	Value or measure
column ^a	30.5 m × 0.5 mm dia
sample size, microliter	0.2
detector	flame ionization
temperature, °C	70
pressure, kPa (psig)	156 (8)
helium flow, mL min	8

^a Stainless steel column coated with *m*-bis(*m*-phenoxyphenoxy)benzene.

Bentone-34 has commonly been used in packed columns (138–139). The retention indices of many benzene homologues on squalane have been determined (140). Gas chromatography of C –C₁₁ aromatic compounds using a Ucon B550X-coated capillary column is discussed in Reference 141. A variety of other separation media have also been used, including phthalic acids (142), liquid crystals (143), and Werner complexes (144). Gel permeation chromatography of alkylbenzenes and the separation of the C₈ aromatics treated with zeolites are described in References 145–148.

Health and Safety Factors

The xylene isomers are flammable liquids and should be stored in approved closed containers with appropriate labels and away from heat and open flames. Limits for transportation by air are 5 L on passenger planes and 60 L on cargo planes.

Flash points and autoignition temperatures are given in Table 11. The vapor can travel along the ground to an ignition source. In the event of fire, foam, carbon dioxide, and dry chemical are preferred extinguishers. The lower and upper explosion limits are 1° o and 7° o.

Table 11. Flash Points and Autoignition Temperatures of the C₈ Aromatic Compounds

Compound	Flash point, °C	Autoignition temperature, °C
PX	27	530
MX	27	530
OX	32	460
EB	15	432

The xylenes are mildly toxic. They are mild skin irritants, and skin protection and the cannister-type masks are recommended. The oral LD₅₀ value for rats is 4300 ppm. The STEL for humans is 150 ppm. Xylenes show only mild toxicity to fish, and the threshold limit for crop damage is 800–2400 ppm. Biodegradation with activated seed is slow, and sewage digestion is impaired by 0.1° o concentrations. In the event of a spill, oil-skimming equipment, adsorbent foam, and charcoal may be used for cleanup.

Uses

The majority of xylenes, which are mostly produced by catalytic reforming or petroleum fractions, are used in motor gasoline (see GASOLINE AND OTHER MOTOR FUELS). The majority of the xylenes that are recovered for petrochemicals use are used to produce PX and OX. PX is the most important commercial isomer. Almost all of the PX is converted to terephthalic acid and dimethylterephthalate, and then to poly(ethylene terephthalate) for ultimate use in fibers, films, and resins.

Almost all of the OX that is recovered is used to produce phthalic anhydride. Phthalic anhydride is a basic building block for plasticizers used in flexible PVC resins, for polyester resins used in glass-reinforced plastics, and for alkyd resins used for surface coatings. OX is also used to manufacture phthalonitrile, which is converted to copper phthalocyanine, a pigment.

Most of the MX that is contained in recovered mixed xylenes is isomerized to PX and OX. MX is also used to produce isophthalic acid and isophthalonitrile. Isophthalic acid is the base of unsaturated polyester resins which have good corrosion resistance, and greater strength and higher modulus than resins derived from phthalic anhydride. Isophthalonitrile is the starting material for the fungicide tetrachloroisophthalonitrile.

Some of the mixed xylenes that are produced are used as solvents in the paints and coatings industry (see SOLVENTS, INDUSTRIAL). However, this use has declined, particularly in the United States as environmental efforts to reduce hydrocarbon emissions into the air have increased.

The EB present in recovered mixed xylenes is largely converted to xylenes or benzene. The EB used to make styrene is predominately manufactured by the alkylation of benzene with ethylene.

BIBLIOGRAPHY

"Xylenes and Ethylbenzene" in *ECT* 1st ed., Vol. 15, pp. 186–194, by M. Lapeyrouse, Esso Research and Engineering Co.; in *ECT* 2nd ed., Vol. 22, pp. 467–507, by H. E. Cier, Esso Research and Engineering Co.; in *ECT* 3rd ed., Vol. 24, pp. 709–744, by D. L. Ransley, Chevron Research Co.

1. *Chemical Economics Handbook*, SRI International, Menlo Park, Calif., 1993.
2. C. L. Yaws, *Chem. Eng.*, 113 (July 21, 1975).
3. C. L. Yaws, *Chem. Eng.*, 73 (Sept. 29, 1975).
4. J. Chao and co-workers, *Hydrocarbon Process.* **59** (8), 117 (1980).
5. J. Chao, *Hydrocarbon Process.* 295 (Nov. 1979).
6. D. R. Stull and co-workers, *The Chemical Thermodynamics of Organic Compounds*, John Wiley & Sons, Inc., New York, 1968.
7. W. L. Nelson, *Oil Gas J.*, 68 (Mar. 19, 1970).
8. D. A. McCauley and A. P. Lien, *J. Am. Chem. Soc.* **74**, 6246 (1950).
9. H. C. Brown and H. Jungk, *J. Am. Chem. Soc.* **77**, 5579 (1955).
10. R. H. Allen and L. D. Yats, *J. Am. Chem. Soc.* **81**, 5289 (1959).
11. *Selected Values of Physical and Thermodynamic Properties of Hydrocarbons and Related Compounds*, API Research Project 44, Carnegie Press, 1953.
12. W. D. Crow and C. Wentrup, *Tetrahedron Lett.* **27**, 3111 (1968).
13. C. J. Egan, *J. Chem. Eng. Data* **5**(3), 298 (1960).

14. S. Ehrenson, *J. Am. Chem. Soc.* **84**, 2681 (1962).

15. A. McCaulay and A. P. Lien, *J. Am. Chem. Soc.* **73**, 2013 (1951).

16. M. Kilpatrick and F. E. Luborsky, *J. Am. Chem. Soc.* **75**, 577 (1953).

17. E. L. Mackor, A. Hofstra, and J. H. van der Waals, *Trans. Faraday Soc.*, 186 (1958).

18. J. Barnard and B. M. Sankey, *Combust. Flame* **12**(4), 345 (1968).

19. L. Errede and M. Swarc, *Q. Rev.* **12**, 301 (1958).

20. L. Errede, R. Gregorian, and J. Hoyt, *J. Am. Chem. Soc.* **82**, 5218 (1960).

21. L. Errede and N. Kroll, *J. Polym. Sci.* **60**, 33 (1962).

22. **U.S. Pat. 3,412,167** (Nov. 19, 1968), J. W. Lewis (to Union Carbide Co.).

23. D. J. Hadley, *Chem. Ind.*, 238 (Feb. 25, 1961).

24. *Hydrocarbon Process.*, 252 (Nov. 1969).

25. *Hydrocarbon Process.*, 238 (Nov. 1981).

26. E. Kamegaya, *Chem. Econ. Eng.* **2**(2), 30 (1970).

27. S. Arika and A. Ohira, *Chem. Econ. Eng.* **5**(7), 39 (1973).

28. N. E. Ockerbloom, *Hydrocarbon Process.*, 101 (Feb. 1972).

29. P. B. de la Mare and J. Robertson, *J. Chem. Soc.*, 279 (1943).

30. F. Condon, *J. Am. Chem. Soc.* **70**, 1963 (1948).

31. **U.S. Pat. 2,282,231** (May 5, 1942), W. J. Mattox (to UOP).

32. O. Maas and J. Russell, *J. Am. Chem. Soc.* **40**, 1561 (1918).

33. O. Maas, E. H. Boomer, and D. M. Morrison, *J. Am. Chem. Soc.* **45**, 1433 (1923).

34. H. C. Brown and H. W. Pearsall, *J. Am. Chem. Soc.* **74**, 191 (1952).

35. W. D. Schaeffer and W. S. Dorsey, *Advances in Petroleum Chemistry and Refining*, Vol. 6, Wiley-Interscience, New York, 1962.

36. W. D. Schaeffer and co-workers, *J. Am. Chem. Soc.* **79**, 5870 (1957).

37. **U.S. Pat. 2,798,103** (July 2, 1957) and **U.S. Pat. 2,951,104** (Aug. 30, 1960), W. D. Schaeffer and J. D. Wordie (to Union Oil Co.).

38. M. J. Minton and co-workers, *J. Phys. Chem.* **71**(11), 3618 (1967).

39. J. Menyhart, *Magy. Kem. Lapja* **22**(1), 18 (1967).

40. F. Casellato, *Erdoel Kohle* **22**, 2 (1969).

41. **E. Ger. Pat. 56,236** (June 6, 1967), W. Jugel.

42. **Brit. Pat. 1,151,606** (May 14, 1969), J. Lumsden (to Imperial Smelting Co.).

43. H. R. Allcock and L. A. Siegel, *J. Am. Chem. Soc.* **86**, 5140 (1964).

44. **U.S. Pat. 3,472,762** (Oct. 14, 1969), A. Goldrup and M. T. Westaway (to British Petroleum).

45. **U.S. Pat. 3,484,500** (Dec. 16, 1969), A. Goldrup, A. B. Morrison, and M. T. Westaway (to British Petroleum).

46. **Brit. Pat. 1,183,524** (Mar. 11, 1970) (to British Petroleum Co.).

47. **U.S. Pat. 3,504,047** (Mar. 31, 1970), J. N. Haresnape (to British Petroleum Co.).

48. **U.S. Pat. 3,456,028** (July 15, 1969), C. G. Gerhold and D. B. Broughton (to UOP).

49. **U.S. Pat. 3,465,055** (Sept. 2, 1969), W. K. T. Gleim, R. C. Wackler, and R. C. Ranquist (to UOP).

50. **U.S. Pat. 2,900,428** (Aug. 18, 1959), C. D. Heaton and W. E. Toland (to Chevron Research Co.).

51. **U.S.S.R. Pat. 176,277** (Nov. 2, 1965), V. S. Bogdanov and co-workers (to Scientific Research Institute for Petrochemical Industries).

52. C. J. Egan and R. V. Luthy, *Ind. Eng. Chem.* **47**, 250 (1955).

53. T. Iwamura, S. Otani, and M. Sato, *Bull. Jpn. Pet. Inst.* **13**(1), 116 (1971).

54. S. Otani, *Chem. Eng.*, 118 (July 20, 1970).

55. *Hydrocarbon Process.*, 133 (Nov. 1977).

56. UOP Petrochemical Technology Conference, 1992, Houston, Tex.

57. J. J. Jeanneret, in R. A. Meyers, ed., *Handbook of Petroleum Refining Processes*, 2nd ed., McGraw-Hill Book Co., Inc., New York, 1997, p. 2.55.

58. J. A. Verdol, *Oil Gas J.*, 63 (June 9, 1969).

59. HRI IFP Petrochemical Seminar, May 8–9, 1995, Houston, Tex.

60. P. Grandio and F. H. Schneider, *Oil Gas J.*, 62 (Nov. 29, 1971).

61. P. Grandio and co-workers, *Hydrocarbon Process.* **51**, 85 (Aug. 1972).

62. S. Han and co-workers, *Oil Gas J.*, 83 (Aug. 21, 1989).

63. Brochure on MTDP-3, Mobil Oil Corp., 1996.

64. F. Gorra, L. Breckridge, R. Sailor, *Oil Gas J.* **90**(41), 60 (Oct. 12, 1992).

65. R. Sailor, presentation at Chiyoda Forum, Saudi Arabia, 1993.

66. Brochure on MSTDP, Mobil Oil Corp., 1996.

67. D. Rothman, *Chem. Week*, 18 (Aug. 30–Sept. 6, 1995).

68. **U.S. Pat. 5,243,117** (Sept. 7, 1993), C. D. Chang and D. S. Shihabi (to Mobil Oil Corp.).

69. Technical data, Mobil Technology Sales and Licensing, Mobil Oil Co., May, 1996.

70. K. Menard, *Oil Gas J.*, 46 (Mar. 16, 1987).

71. **U.S. Pat. 3,467,724** (Sept. 16, 1969), S. A. Laurich (to Chevron Research Co.).

72. *Hydrocarbon Process.*, 253 (Nov. 1979).

73. **U.S. Pat. 3,177,265** (Apr. 6, 1965), G. C. Lammens (to Standard Oil Co., Indiana).

74. **Belg. Pat. 617,795** (Sept. 14, 1962), (to Standard Oil Co., Indiana).

75. R. J. Desiderio and co-workers, *Hydrocarbon Process.*, **53**(8), 81 (1974).

76. *Hydrocarbon Process.*, 239 (1981).

77. Y. Hatanaka and T. Nakamura, *Oil Gas J.* **70**(47), 60 (1972).

78. *Hydrocarbon Process.*, 254 (Nov. 1979).

79. D. L. McKay and H. W. Goard, *Chem. Eng. Prog.* **61**(11), 99 (1966).

80. D. L. McKay and co-workers, *Chem. Eng. Prog.* **62**(11), 104 (1966).

81. *Hydrocarbon Process.*, 132 (Mar. 1995).

82. G. Parkinson, *Chem. Eng.* **103**(9), 23 (Sept. 1996).

83. N. P. Wynn, *Chem. Eng. Prog.* **88**(3), 52 (1992).

84. D. B. Broughton and co-workers, *The Separation of p-Xylene from C₈ Hydrocarbon Mixtures by Parex Process*, AIChE, Puerto Rico, May 1970.

85. D. B. Broughton and co-workers, *Chem. Eng. Prog.* **66** (9), 70 (1970).

86. D. P. Thornton, Jr., *Hydrocarbon Process.*, 151, (Nov. 1970).

87. D. B. Broughton, *Chem. Eng. Prog.* **73**(10), 49 (1977).

88. *Hydrocarbon Process.*, 256 (Nov. 1979).

89. **U.S. Pat. 3,663,638** (May 16, 1972), R. W. Neuzil (to UOP).

90. **U.S. Pat. 3,665,046** (May 23, 1972), A. J. DeRosset (to UOP).

91. **U.S. Pat. 3,686,342** (Aug. 22, 1972), R. W. Neuzil (to UOP).

92. U.S. Pat. 3,636,180 (Jan. 18, 1972), D. B. Broughton (to UOP).

93. U.S. Pat. 3,696,107 (Oct. 3, 1972), R. W. Neuzil (to UOP).

94. Ref. 57, p. 2.45.

95. S. Otani and co-workers, *Chem. Econ. Eng. Rev.*, 56 (June 1971).

96. S. Otani, *Chem. Eng.*, 106 (Sept. 17, 1973).

97. S. Otani and co-workers, *Jpn. Pet. Inst. Bull.* **16**(1), 60 (1974).

98. U.S. Pat. 3,761,533 (Sept. 25, 1973), S. Otani and co-workers (to Toray Industries, Inc.).

99. G. Ash and co-workers, *Rev. de IFP* **49**(5), 541 (Sept.–Oct. 1994).

100. M. Seko, *Oil Gas J.*, 81 (July, 1979).

101. M. Seko, T. Miyake, and K. Inada, *IEC Prod. Res. Dev.* **18** (4), 263 (1979).

102. M. Seko, T. Miyake, and K. Inada, *Hydrocarbon Process.*, 133 (Jan. 19, 1980).

103. M. Seko, H. Takeuchi, and T. Inada, *IEC Prod. Res. Dev.* **21**, 656 (1982).

104. Y. Igarashi and T. Ueno, *A New Xylene Separation Process*, ACS Meeting, Atlantic City, N.J., 1968.

105. *Hydrocarbon Process.*, 254 (Nov. 1969).

106. Y. Igarashi, *Jpn. Chem. Q.* **IV**(4), 27 (1969).

107. T. Ueno, *Bull. Jpn. Pet. Inst.* **12** (May 1970).

108. G. R. Herrin and E. H. Martel, *Pet. Petrochem. Int.* **12**(7), 74 (1972).

109. J. C. Davis, *Chem. Eng.*, 77 (Aug. 9, 1971).

110. T. Ueno and T. Nakana, *Eighth World Pet. Congr. Proc.* **4**, 187 (1971).

111. S. Atriki and A. Ohira, *Chem. Econ. Eng. Rev.* **5**(7), 30 (1973).

112. J. J. H. Masseling, *Chem. Tech.*, 714 (Nov. 1976).

113. *Hydrocarbon Process.*, 255 (Nov. 1979).

114. H. F. Uhlig and W. C. Pfefferle, *Am. Chem. Soc. Div. Pet. Chem. D*, 154 (Sept. 1969).

115. P. M. Pitts, Jr., J. E. Connor, Jr., and L. N. Leum, *Ind. Eng. Chem.* **47**(4), 770 (1955).

116. U.S. Pat. 2,976,332 (Mar. 21, 1961), L. N. Leum and J. E. Connor, Jr. (to Atlantic Refining Co.).

117. C. V. Berger, *Hydrocarbon Process.*, 173 (Sept. 1973).

118. *Hydrocarbon Process.*, 215 (Mar. 1993).

119. Ref. 57, p. 2.37.

120. S. Otani and co-workers, *Chem. Econ. Eng. Rev.*, 56 (June 1971).

121. S. Otani, *Chem. Eng.*, 106 (Sept. 17, 1973).

122. Y. Tsuchiya and co-workers, *Jpn. Pet. Inst. Bull.* **16**(1), 60 (1974).

123. U.S. Pat. 4,101,596 (July 18, 1978), K. M. Mitchell and J. J. Wise (to Mobil Oil Corp.).

124. Brochure on MHAI and MLPI, Mobil Oil Corp., 1996.

125. U.S. Pat. 3,856,872 (Dec. 24, 1974), R. A. Morrison (to Mobil Oil Corp.).

126. N. Y. Chen, W. E. Garwood, F. G. Dwyer, *Shape Selective Catalysis in Industrial Applications*, Marcel Dekker, Inc., New York, 1989.

127. U.S. Pat. 4,899,011, (Feb. 6, 1990), Chu and co-workers (to Mobil Oil Corp.).

128. Y. Y. Huang and co-workers, **30d**, 1991 AIChE Spring National Meeting, Houston, Apr. 7–11, 1991.

129. Amelse and co-workers, **33a**, 1988 AIChE Summer National Meeting, Denver, Aug. 21–24, 1988.

130. H. C. Ries, *Process Economics Program 25A*, Stanford Research Institute, Menlo Park, California, July 1970, pp. 41–53.

131. *Hydrocarbon Process.*, 240 (Nov. 1981).

132. *Hydrocarbon Process.*, 109 (Aug. 1969).

133. *Hydrocarbon Process.*, 253 (Nov. 1969).

134. U.S. Pat. 3,793,384 (Feb. 19, 1974), J. G. Chenoweth and co-workers (to Imperial Chemical Industries).

135. U.S. Pat. 3,860,668 (Jan. 14, 1975), J. K. January and A. Marchant (to Imperial Chemical Industries).

136. *World Petrochemicals*, SRI International, Menlo Park, Calif., 1996.

137. *Chemical Economics Handbook*, SRI International, Menlo Park, Calif., 1996.

138. J. M. Vergnaud, *J. Chromatogr.* **27**, 54 (1967).

139. C. F. Raley and J. W. Kaufman, *Anal. Chem.* **40**(8), 1371 (1968).

140. M. Dimov and D. Papzova, *J. Chromatogr.* **137**, 265 (1977).

141. F. Baumann and S. M. Csicsery, *J. Chromatogr.* **26**, 262 (1967).

142. M. F. Burke and L. B. Rogers, *J. Chromatogr. Sci.* **6**(1), 75 (1968).

143. A. B. Richmond, *J. Chromatogr. Sci.* **9**(9), 571 (1971).

144. A. C. Bhattacharyya and A. Bhattacharjee, *J. Chromatogr.* **41**(3–4), 446 (1969).

145. M. Popl, J. Coupek, and S. Pokorny, *J. Chromatogr.* **104**, 135 (1975).

146. U.S. Pat. 3,653,184 (Apr. 4, 1972), B. M. Drinkard, P. T. Allen, and E. H. Unger (to Mobil Oil Corp.).

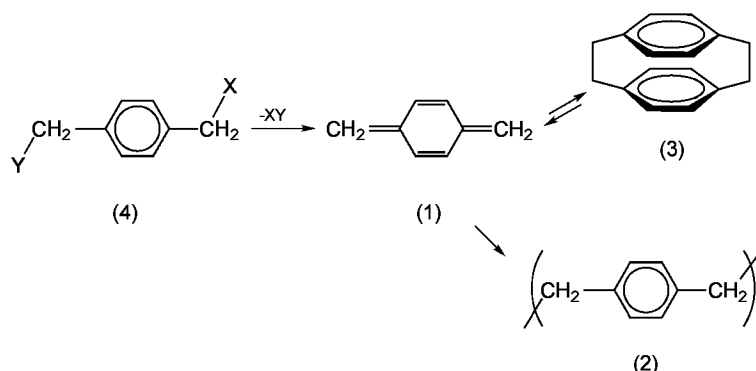
147. U.S. Pat. 3,656,278 (Apr. 18, 1972), B. M. Drinkard, P. T. Allen, and E. H. Unger (to Mobil Oil Corp.).

148. U.S. Pat. 3,724,170 (Apr. 3, 1973), P. T. Allen, B. M. Drinkard, and E. H. Unger (to Mobil Oil Corp.).

William J. Cannella
Chevron Research & Technology Company

XYLYLENE POLYMERS

In a process capable of producing pinhole-free coatings of outstanding conformality and thickness uniformity through the unique chemistry of *p*-xylylene (PX) [502-86-3] (1), a substrate is simply exposed to a controlled atmosphere of pure gaseous monomer. The coating process is best described as a vapor deposition polymerization (VDP). The monomer molecule is thermally stable, but kinetically very reactive toward polymerization with other molecules of its kind. Although it is stable as a rarified gas, upon condensation it polymerizes spontaneously to produce a coating of high molecular weight, linear poly(*p*-xylylene) (PPX) [25722-33-2] (2). This article emphasizes recent VDP developments. There have been several reviews of the subject (1), which offer a more thorough treatment of early developments in the field.



In the commercial Gorham process (2), the extremely reactive PX is conveniently generated by the thermal cleavage of its stable dimer, *cyclo*-di-*p*-xylylene (DPX), a [2.2]paracyclophane [1633-22-3] (3). In many instances, substituents attached to the paracyclophane framework are carried through the process unchanged, ultimately becoming substituents of the polymer in the coating.

The process takes place in two stages that must be physically separate but temporally adjacent. Figure 1 presents a schematic of a typical parylene deposition process, also indicating the approximate operating conditions.

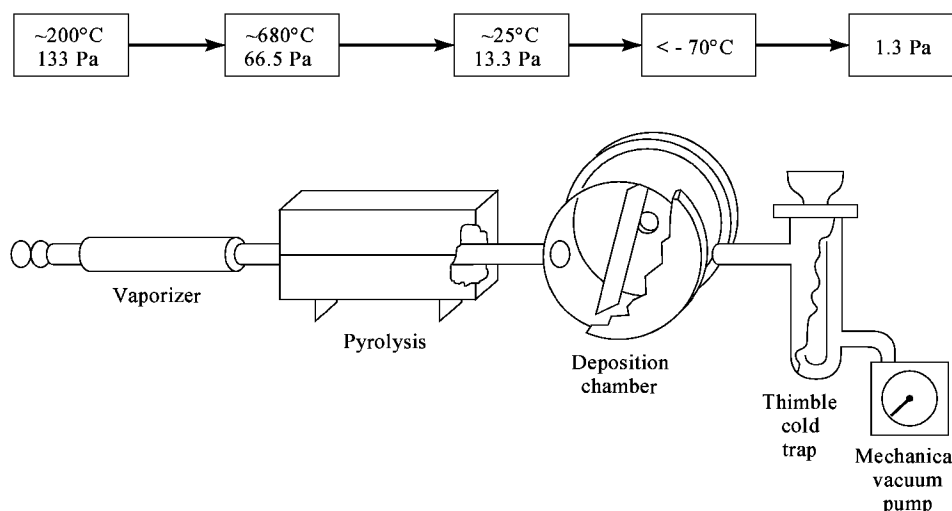


Fig. 1. Parylene deposition apparatus. To convert Pa to torr, multiply by 0.0075.

The PPXs formed as coatings in the Gorham process are referred to generically as the parylenes. The terms Parylene N [25722-33-2], Parylene C [9052-19-1], and Parylene D [52261-45-7] refer specifically to polymers produced as coatings by the Gorham process using the dimers DPXN, DPXC [28804-46-8], and DPXD [30501-29-2], respectively, originally marketed by Union Carbide Corporation.

The parylene process has certain similarities with vacuum metallizing. The principal distinction is that truly conformal parylene coatings are deposited even on complex, three-dimensional substrates, including on sharp points and in hidden or recessed areas. Vacuum metallizing, on the other hand, is a line-of-sight coating technology. Whatever areas of the substrate cannot be "seen" by the evaporation source are "shadowed" and remain uncoated (see VACUUM TECHNOLOGY).

The *p*-xylylene species plays a central role in the coating process itself as well as in the making of the dimers which are used as feedstocks for the coating process. Polymers and dimers have both been made from precursor *p*-xylene compounds (4) featuring a variety of X and Y leaving groups. The conditions of the reaction determine the relative amounts of the resulting dimeric or polymeric products. Dilution is of course the key element in any procedure which offers a high yield of dimer.

The modest commercial success the *p*-xylylene dimer-based Gorham process has achieved to date is readily attributed to the fact that thermal cleavage of cyclic dimer produces the *p*-xylylene monomer in essentially quantitative yield, while at the same time producing no gaseous by-products. In a gas-to-solid coating process, any gaseous entities generated from the leaving groups X and Y, necessarily formed in volumes comparable to the volume of the monomeric *p*-xylylene generated, would at the very least need to be exhausted through the pumping system, thereby slowing the process down. Moreover, certain of the most effective leaving groups XY, such as halogens or halogenide acids, would create a corrosion hazard both for any sensitive substrates to be coated and for the deposition equipment itself.

Gorham Process Monomers

The eight-carbon monomer PX is generated in the first stage of the parylene process by heating gaseous dimer as it passes through a high temperature

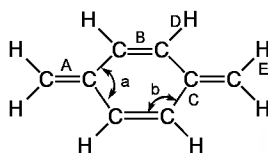


Fig. 2. Structure of PX monomer molecule from electron diffraction (8). Bond lengths: C—C, A—B = 0.1381 ± 0.0008 nm; C—C, C—H, D—E = 0.1451 ± 0.0007 nm; C—H, D—E = 0.1116 ± 0.0035 nm. Bond angles: a = $122.2 \pm 3.7^\circ$, b = $118.9 \pm 1.9^\circ$.

By trapping PX at liquid nitrogen temperature and transferring it to THF at -80°C , the ^1H nmr spectrum could be observed (9). It consists of two sharp peaks of equal area at chemical shifts of 5.10 and 6.49 ppm downfield from tetramethylsilane (TMS). The fact that any sharp peaks are observed at all attests to the absence of any significant concentration of unpaired electron spins, such as those that would be contributed by the biradical (11). Furthermore, the chemical shift of the ring protons, 6.49 ppm, is well upfield from the typical aromatic range and more characteristic of an olefinic proton. Thus the olefin structure (1) for PX is also supported by nmr.

A particularly useful property of the PX monomer is its enthalpy of formation. Conventional means of obtaining this value, such as through its heat of combustion, are, of course, excluded by its reactivity. An experimental attempt was made to obtain this measure of chemical reactivity with the help of ion cyclotron resonance; a value of 209 ± 17 kJ/mol (50 ± 4 kcal/mol) was obtained (10). Unfortunately, the technique suffers from lack of resolution in addition to experimental imprecision. It is perhaps better to rely on molecular orbital calculations for the formation enthalpy. Using a semiempirical molecular orbital technique, which is tuned to give good values for heat of formation on experimentally accessible compounds, the heat of formation of *p*-xylene has been computed to be 234.8 kJ/mol (56.1 kcal/mol) (11).

Successful *p*-Xylylene VDP Monomers. Within the limits mentioned above, it is frequently possible, and often desirable, to modify the *p*-xylylene monomer by attaching to it certain substituents. Limitations on such modifications lie in three areas: reactivity, performance in the coater, and cost.

Reactivity. Although the reactivity which enables the gas–solid polymerization to proceed is a characteristic of the eight-carbon *p*-xylylene tetraolefin system, it is possible to subdue that reactivity. For example, by attaching electron-withdrawing substituents to the alpha positions and thereby further delocalizing the π -electrons of the highly reactive *p*-xylylene nucleus, it is in several instances possible to prepare *p*-xylenes that are so stable that they can be isolated and handled as normal organic compounds (Fig. 3). These sorts of substitutions must of course be avoided if the goal is to make polymer.

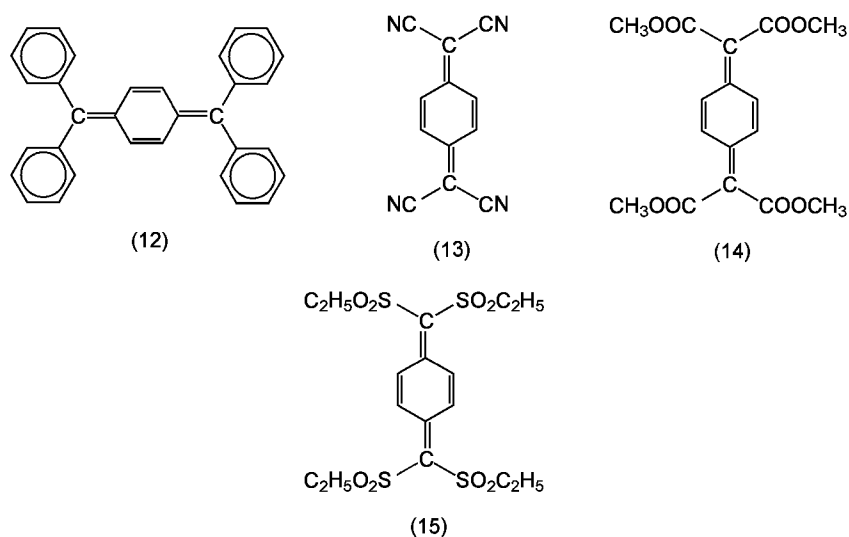


Fig. 3. Isolatable *p*-xylylene derivatives: (12), Thiele's hydrocarbon - 1904 [26392-12-1]; (13), tetracyanoquinodimethane [1518-16-7] (TCNQ); (14), tetrakis(methoxycarbonyl)-quinodimethane [65649-20-9]; (15), tetrakis(ethylsulfonyl)quinodimethane [84928-90-5].

It is also possible to interfere with the polymerization by attaching at the alpha positions either too many groups, or groups which are too bulky. Four chlorine atoms (12) or four methyl groups (13) seem to be sufficient to hinder the production of polymer. These crowded *p*-xylylene monomers can be polymerized, but not through a VDP process.

Thus, except for electron-withdrawing or bulky substituents, at least from the standpoint of reactivity toward polymerization, modification by most other substituents is possible.

Performance in Coater. The modified monomer should perform well in commercial deposition equipment. Performance considerations include the growth rate of the coating, the uniformity of thickness of the coating over the chamber volume, and the efficiency with which the dimer is converted to useful coatings on the substrates.

An important further constraint is the fact that economic considerations in the construction of deposition equipment normally lead to a preference for an ambient temperature deposition chamber. Control of deposition temperature is possible, but it adds both equipment expense and operational complexity.

The vapor pressure of a parylene monomer is a prime factor in determining how rapidly a coating grows when exposed to an atmosphere of monomer at a given pressure. Vapor pressure is reduced as molecular weight increases, thereby increasing the monomer's tendency to condense and, along with it, increasing the VDP growth rate. The presence of polar functionality in the molecule further depresses vapor pressure. But too low a vapor pressure makes it difficult to transport gaseous monomer from point to point in the deposition chamber. Hence, some optimum value of monomer volatility is expected.

The widely used Parylene C owes its popularity principally to the room temperature volatility of its monomer. The Parylene C monomer, chloro-*p*-xylylene, has become the de facto performance standard. By comparison, the Parylene N monomer, *p*-xylylene itself, is too volatile and would perform better in a sub-ambient temperature deposition system. The Parylene D monomer, dichloro-*p*-xylylene [85586-88-5] is too heavy, and causes distribution problems in larger deposition systems.

Cost. It is necessary to produce the feedstock from which the monomer is generated, viz, the dimer, at a cost which can be supported by the commercial application, and yet allow it to be economically competitive with all other alternative ways to achieve the same end result. This factor often, but not always, seriously limits the amount of effort that can be put into dimer synthesis and purification.

Other, Related Processes

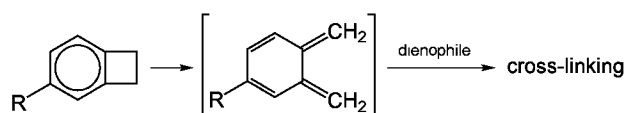
VDP processes using means other than the pyrolytic cleavage of DPX (Gorham process) to generate the reactive monomer are also known, although none are practiced commercially at the time of this writing (ca 1997).

Photopolymerization and Plasma Polymerization. The use of ultraviolet light alone (14) as well as the use of electrically excited plasmas or glow discharges to generate monomers capable of undergoing VDP have been explored. The products of these two processes, called plasma polymers, continue to receive considerable scientific attention. Interest in these approaches is enhanced by the fact that the feedstock material from which the monomer capable of VDP is generated is often inexpensive and readily available. In spite of these widespread scientific efforts, however, commercial use of the technologies is quite limited.

When *p*-xylene is used as the monomer feed in a plasma polymer process, PX may play an important role in the formation of the plasma polymer. The plasma polymer from *p*-xylene closely resembles the Gorham process polymer in the infrared, although its spectrum contains evidence for minor amounts of nonlinear, branched, and cross-linked chains as well. Furthermore, its solubility and low softening temperature suggest a material of very low molecular weight (15).

VDP Polyimides. Polyimide films have also been prepared by a kind of VDP (16). The poly(amic acid) layer is first formed by the coevaporation and condensation of two monomers, followed by copolymerization on the substrate. The imidization is carried out in a separate baking step (see POLYIMIDES).

***o*-Xylylene/BCB.** Thermosetting resins based on benzocyclobutene (BCB)



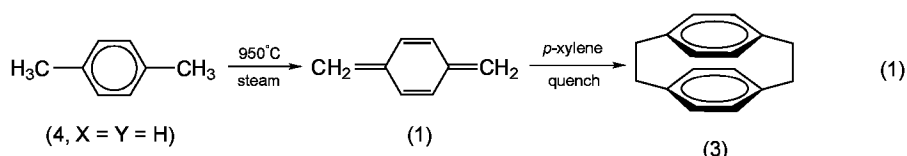
chemistry have been reported (17). In these condensed phase cures, the *ortho*-xylylene isomer is the key reactive intermediate. From the behavior of this energetically similar *ortho* isomer, the value of the para configuration's rendering any ring closure reaction, analogous to cyclobutene formation from the *ortho* isomer, geometrically forbidden can be appreciated.

Dimer

In contrast to the extreme reactivity of the monomeric PX (1) generated from it, the dimer DPX (3) feedstock for the parylene process is an exceptionally stable compound. Because of their chemical inertness, dimers in general do not exhibit shelf-life limitations. Although a variety of substituted dimers are known in the literature, at present only three are commercially available: DPXN, DPXC, and DPXD, which give rise to Parylene N, Parylene C, and Parylene D, respectively.

The unsubstituted C-16 hydrocarbon, [2.2]paracyclophane (3), is DPXN. Both DPXC and DPXD are prepared from DPXN by aromatic chlorination and differ only in the extent of chlorination; DPXC has an average of one chlorine atom per aromatic ring and DPXD has an average of two.

Manufacture. For the commercial production of DPXN (di-*p*-xylylene) (3), two principal synthetic routes have been used: the direct pyrolysis of *p*-xylene (4, X = Y = H) and the 1,6-Hofmann elimination of ammonium (HNR₃⁺) from a quaternary ammonium hydroxide (4, X = H, Y = NR₃⁺). Most of the routes to DPX share a common strategy: PX is generated at a controlled rate in a dilute medium, so that its conversion to dimer is favored over the conversion to polymer. The polymer by-product is of no value because it can neither be recycled nor processed into a commercially useful form. Its formation is minimized by careful attention to process engineering. The chemistry of the direct pyrolysis route is shown in equation 1:



First, *p*-xylene is dehydrogenated pyrolytically in the presence of steam at about 950°C to give *p*-xylylene (PX), which in turn forms di-*p*-xylylene (DPX) when quenched in liquid xylene. The xylene is recycled to the pyrolysis vessel. Yields and conversion efficiency are satisfactory. However, several engineering challenges need to be overcome, including the choice of a suitable diluent; establishing optimal residence time, vapor velocity, and operating pressure during pyrolysis; and the design and construction of novel equipment to withstand the highly corrosive reaction environment.

The Hofmann elimination route, of which many versions exist, can be carried out at much lower temperatures in conventional equipment. The PX is generated by a 1,6-Hofmann elimination of amine from a quaternary ammonium hydroxide in the presence of a base. This route gives yields of 17–19%. Undesired polymeric products can be as high as 80% of the product. In the presence of a polymerization inhibitor, such as phenothiazine, DPXN yields can be increased to 50%.

In the 1,6-elimination of *p*-trimethylsilylmethylbenzyltrimethylammonium iodide with tetrabutylammonium fluoride, yields as high as 56% have been reported (18). The starting materials are not readily accessible, however, and are costly.

The yield can be raised to 28% if the Hofmann elimination is conducted in the presence of a water-soluble copper or iron compound (19). Further improvements up to 50% were reported when the elimination was carried out in the presence of ketone compounds (20). Further beneficial effects have been found with certain cosolvents, with reported yields of greater than 70% (8).

DPXC and DPXD. The economic pressure to control dimer costs has had an important effect on what is in use today (ca 1997). Attaching substituents to the ring positions of a [2.2]paracyclophane does not proceed with isomeric exclusivity. Indeed, isomeric purity in the dimer is not an essential requirement for the obtaining of isomeric purity, eg, monosubstituted monomer, in the pyrolysis. Any mixture of the four possible heteronuclearly disubstituted dichloro[2.2]paracyclophanes, will, after all, if pyrolyzed produce the same monomer molecule, chloro-*p*-xylylene [10366-09-3] (16) (Fig. 4).

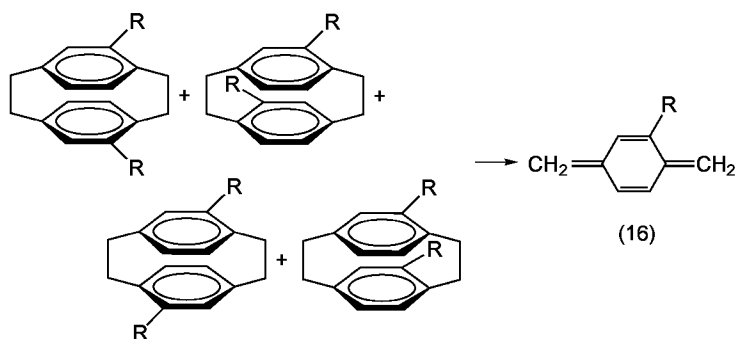


Fig. 4. Isomeric dichloro[2.2]paracyclophanes produce the same xylylene.

Although DPXC and DPXD prepared by the chlorination of DPXN are relatively complex mixtures, after pyrolytic cleavage the resulting mixture

of monomers is considerably simpler. Thus DPXC, when pyrolyzed, gives predominantly monochloro PX, which is accompanied by small but significant amounts of PX and dichloro PX. The resulting polymer, Parylene C, consequently has an average of about one chlorine atom per repeat unit. However, it contains significant amounts of unchlorinated, as well as dichlorinated, repeat units.

DPXC and DPXD are prepared from DPXN by chlorinating to different extents. The conditions are controlled to favor aromatic ring chlorination to the exclusion of the free-radical chlorination of the ethylene bridges. However, the chlorination products are complex mixtures of the homologues DPXN, monochloro DPX, dichloro DPX, trichloro DPX, and tetrachloro DPX, and even higher homologues, as well as the several possible isomers of each.

New synthetic routes for the preparation of homologically pure dichloro DPX and tetrachloro DPX have been reported through the 1,6-Hofmann elimination of chlorinated *p*-methylbenzyltrimethylammonium hydroxide. In the case of dichloro DPX, yields of 30% were reported (21). In the presence of ketone compounds, yields were increased to 50% (20).

Purification. Unsubstituted di-*p*-xylylene (DPXN) is readily purified by recrystallization from xylene. It is a colorless, highly crystalline solid. The principal impurity is polymer, which fortunately is insoluble in the recrystallization solvent and easily removed by hot filtration.

In purifying DPXC and DPXD, care must be taken not to disturb the homologue composition, so that product uniformity is maintained. For example, a recrystallization of DPXC from ethanol would give a higher melting, more crystalline dimer material, at the expense of a decrease in yield owing to the removal of otherwise useful isomers, but the polymer made from it would not be identical to the historical Parylene C, as defined by its preparation from the chlorination mixture. The real purification issues are the removal of insoluble residues and any components that contain aliphatic side-chain chlorine. Although ring-substituted chlorine is stable, side-chain chlorine can give rise to hydrogen chloride gas under the conditions of the parylene process, or subsequent to it, which in certain applications could initiate substrate corrosion. Fortunately, the aliphatic chlorine problem can be minimized by proper attention to process detail.

Properties. The DPXs are all crystalline solids; melting points and densities are given in Table 1. Their solubility in aromatic hydrocarbons is limited. At 140°C, the solubility of DPXN in xylene is only about 10%. DPXC is more readily soluble in chlorinated solvents, eg, in methylene chloride at 25°C its solubility is 10%. In contrast, the corresponding figure for DPXN is 1.5%.

Table 1. Properties of Parylene Dimers

Dimer	Melting point, °C	Density, g cm ³
DPXN	284 ^a	1.22
DPXC	140–160 ^b	1.30
DPXD	170–195 ^b	1.45

^a Decomposes.
^b Mixture of homologues and their isomers.

The structure of DPXN was determined in 1953 from x-ray diffraction studies (22). There is considerable strain energy in the buckled aromatic rings and distorted bond angles. The strain has been experimentally quantified at 130 kJ mol (31 kcal mol) by careful determination of the formation enthalpy through heat of combustion measurements (23). The release of this strain energy is doubtless the principal reason for success in the particularly convenient preparation of monomer in the parylene process.

Polymer

The linear polymer of PX, poly(*p*-xylylene) (PPX) (2), is formed as a VDP coating in the parylene process. The energetics of the polymerization set it apart from all other known polymerizations and enable it to proceed as a vapor deposition polymerization.

Thermodynamic Considerations.

On the basis of the value for the enthalpy of formation of *p*-xylylene, Δ*H*_f⁰(PX), the enthalpy of polymerization, Δ*H*_{polym}⁰ = Δ*H*_f⁰(PPX) − Δ*H*_f⁰(PX), can be estimated. No experimental combustion data are available for high molecular weight poly(*p*-xylylene) as it is formed in the parylene process, Δ*H*_f⁰(PPX).

For crystalline [2,2]paracyclophane [(1), DPXN], a Δ*H*_f⁰ of +154.4 kJ mol (+36.9 kcal mol) is reported (23). The hypothetical transformation of crystalline DPXN into polymer is accompanied by the release of 129.7 kJ mol (31.0 kcal mol) of paracyclophane strain energy per mole of paracyclophane, and 12.6 kJ mol (3.0 kcal mol) per polymer repeat unit as a result of the bibenzyl hyperconjugative stabilization, which is permitted in the polymer but excluded by geometry in the dimer. Thus the standard enthalpy of formation for the hypothetical 100% crystalline poly(*p*-xylylene) is estimated to be −0.3 kJ mol (−0.05 kcal mol), assuming that the energies associated with crystallinity are the same in both cases. Although it might be acceptable to assume that such energies per repeat unit are similar in the crystalline polymer and crystalline dimer, Parylene N, as produced by the parylene process, is typically only about 57% crystalline. Using a value of 14.1 kJ mol (3.37 kcal mol) for the heat of fusion for poly(*p*-xylylene) (24), the standard formation enthalpy for Parylene N, as it is typically deposited in the parylene process, Δ*H*_f⁰(Parylene N), is +5.7 kJ mol (+1.4 kcal mol).

In estimating the enthalpy of polymerization, the physical state of both starting monomer and polymer must be specified. Changes in state are accompanied by ethalpy changes. Therefore, they also affect the level of the polymerization enthalpy. The Δ*H*_f⁰ for *p*-xylylene previously mentioned is applicable to the monomer as an ideal gas. To make comparisons with other polymerization processes, most of which start with condensed monomer, a heat of vaporization for *p*-xylylene is needed. It is assumed herein that it is the same as that for *p*-xylene, 42.4 kJ mol (10.1 kcal mol). Thus the Δ*H*_f⁰ of the liquid monomer *p*-xylylene is 192.3 kJ mol (46.0 kcal mol).

The enthalpy of polymerization of unannealed (57% crystalline) Parylene N, as it is deposited, starting with liquid monomer, Δ*H*_{polym(lu)}⁰, is −186.6 kJ mol (−44.6 kcal mol). This is an exceptionally high value compared with those of other addition polymers, which generally fall in the −60 to −100 kJ mol (−14.3 to −23.8 kcal mol) range. It quantifies the vigor of the polymerization. Because the source of polymerization enthalpy is within the *p*-xylylene system, substituents affect it only to a minor extent. All parylenes are expected to have a similar molar enthalpy of polymerization. An experimental value for the heat of polymerization of Parylene C has appeared. Using the gas evolution from the liquid nitrogen cold trap to measure thermal input from the polymer, and taking advantage of a peculiarity of Parylene C at −196°C to polymerize abruptly, perhaps owing to the arrival of a free radical, a Δ*H*_{polym}⁰ = 152 ± 8 kJ mol (36.4 ± 2.0 kcal mol) at −196°C was reported (25). The correction from −196°C to room temperature is estimated at −17 kJ mol, bringing this experimental value for Parylene C closer to the calculated value for Parylene N. It is assumed that *S*_{polym}⁰ is 0 at 0 K (3rd law), 125 J (mol·K) [30 cal (mol·K)] at 298 K, and proportional to *T* in between, a crude assumption, but appropriate to the current level of knowledge. Thus experiment and calculation are in harmony in quantifying the exceptional exothermicity of parylene polymerization.

The thermodynamic ceiling temperature (26) *T*_c for a polymerization is computed by dividing the Δ*H*_{polym}⁰ by the standard entropy of polymerization, Δ*S*_{polym}⁰. The *T*_c is the temperature at which monomer and polymer are in equilibrium in their standard states at 25°C (298.15 K) and 101.3 kPa (1 atm). (In the case of *p*-xylylene, such a state is, of course, purely hypothetical.) The *T*_c quantifies the binding forces between monomer units in a polymer and measures the tendency of the polymer to revert back to monomer. In other systems, the *T*_c indicates a temperature above which the polymer is unstable with respect to its monomer, but in the case of parylene it serves rather as a means of comparing the relative stability of the polymer with

respect to its reversion to monomer. For computing the T_g , however, the standard entropies of polymerization are required.

The standard polymerization entropies can be estimated from the following. The standard entropy S° for PX as an ideal gas is computed by a group-contribution method (27) to be 310.6 J (mol·K) [74.24 cal (mol·K)]. The entropy of vaporization for PX is assumed to be the same as that of *p*-xylene, 104.7 J (mol·K) [25.03 cal (mol·K)] (28). Therefore, the S° for liquid PX is 205.9 J (mol·K) [49.21 cal (mol·K)]. Noting that the experimental specific heat C_p of PPX follows that of polystyrene over the range of 160 to 340 K (29), it can be assumed that the proportionality continues down to 0 K and that the factor 135–116 at 298 K can be applied to the known S° for polystyrene [S° = 128.5 J (mol·K) or 30.70 cal (mol·K)] (30). It follows that the S° for as-deposited 57° crystalline Parylene N is 149.5 J (mol·K) [35.73 cal (mol·K)]. Therefore, $\Delta S^\circ_{\text{polym} \rightarrow \text{g}}$ = 161.1 J (mol·K) [38.50 cal (mol·K)] and $\Delta S^\circ_{\text{polym} \rightarrow \text{l}}$ = 56.4 J (mol·K) [13.48 cal (mol·K)].

The results of the above polymerization thermodynamics calculations for parylene are compared to similar data for typical addition polymers in Table 2. The T_g quantifies the stability of the polymer only with respect to reversion to monomer. When PPX is thermally degraded (ca 500°C), a mixture of degradation products including hydrogen gas, *p*-xylene, toluene, and *p*-methylstyrene is observed (31), suggesting that the path taken in thermal degradation requires the cleavage of bonds other than those formed in the polymerization, very likely starting with the methylene C–H bond. Complete replacement of the methylene hydrogens in PPX with fluorine gives a polymer with substantially better stability at elevated temperatures (32).

Table 2. Entropies, Enthalpies, and Ceiling Temperatures for the Polymerization of Various Monomers at 25°C (298.15 K) and 101.3 kPa (1 atm)^a

Monomer	Liquid			Gas		
	ΔH° , kJ mol ^b	ΔS° , J (mol·K) ^b	T_g , °C	ΔH° , kJ mol ^b	ΔS° , J (mol·K) ^b	T_g , °C
ethylene	108.4	173.6	351			
propylene	81.6	116.3	429			
isoprene	74.9	101.3	467	101.3	187.0	268
styrene	69.9	104.6	395	113.4	212.1	262
methyl meth-acrylate	55.2	117.2	198			
α-methylstyrene	35.1	103.8	65			
<i>p</i> -xylylene	186.6	56.4	3035	229.1	161.1	1149

^a Ref. 26.
^b To convert J to cal, divide by 4.184.

The enthalpy liberated on the VDP of parylene is real and in an adiabatic situation causes a rise in temperature of the coated substrate. For Parylene C, 229.1 kJ mol (54.7 cal mol) corresponds to 1654 J g (395 cal g) whereas its specific heat at 25°C is only 1.00 J (g·K) [0.239 cal (g·K)] (33). In most practical situations, however, the mass of parylene deposited is dwarfed by the substrate mass, and the heat of polymerization is dissipated within the coated substrate over the time required to deposit the coating with minimal actual temperature rise.

Polymerization Mechanism. The physical processes of condensation and diffusion must be considered along with the *p*-xylylene polymerization chemistry for a proper understanding of what happens microscopically during vapor deposition polymerization (34). These processes point to an important distinction between VDP and vacuum metallization, ie, that in the latter, adsorption is followed by a surface reorganization of the existing deposited material, and diffusion of incoming species through the bulk is nonexistent. In most parylene depositions, a coating forms from gaseous monomer under steady-state conditions.

Gaseous monomer is transported to the location within the coating where it is to be consumed to produce polymer by an initial condensation, followed by diffusion. The net flux of monomer molecules through the growth interface, ie, the outer boundary of the coating, between the gaseous and condensed phases, needed to sustain growth at a given rate can be readily calculated [for Parylene C, 10 μm h requires 1.55 × 10¹⁵ (cm²·s)]. Comparing a net flux so obtained with the flux of molecules that according to the kinetic theory of gases are striking the growth surface [$Z = PN_0/\sqrt{2\pi MRT}$] for the conditions typical of parylene deposition, a large difference (two or three orders of magnitude) is observed. For Parylene C monomer at a pressure of 1.3 Pa (10 μm Hg) and 25°C, $Z = 6.7 \times 10^{17}$ (cm²·s). For each molecule that eventually enters the coating, some hundred or thousand molecules strike the growth interface. Those that condense and do not react must, of course, evaporate. The term "sticking coefficient" has sometimes been borrowed from vacuum metallization to describe this ratio of incident molecules to consumed molecules. However, the VDP situation is not adequately described by hard spheres bouncing off a growth interface. Every incident molecule spends at least some time in the polymeric coating phase beyond the growth interface before it is lost again to the gas phase.

Because most of the condensing molecules evaporate, condensation equilibrium at the growth surface can be assumed, to a good approximation. The concentration of monomer dissolved in the coating near the growth interface is, therefore, governed by Henry's law, and monomer concentration in polymer solution increases proportionately to the partial pressure of monomer in the gas phase. Furthermore, as the temperature is lowered, or as higher molecular weight monomers of lower volatility are selected, monomer concentration at the growth interface increases. In most practical situations, these Henry's law effects dominate in determining growth rates for VDP coatings by regulating monomer concentration within the coating. For each monomer, there exists a threshold condensation temperature, T_{tc} , above which the rate of growth of coating is, for all practical purposes, zero (Table 3), but this phenomenon is governed by the competition between initiation and propagation chemistries, discussed herein.

Table 3. Threshold Condensation Temperatures T_{tc} for Substituted *p*-Xylylene Monomers

Monomer	T_{tc} , °C
<i>p</i> -xylylene	30
2-methyl- <i>p</i> -xylylene	60
2-ethyl- <i>p</i> -xylylene	90
2-chloro- <i>p</i> -xylylene	90
2-acetyl- <i>p</i> -xylylene	130
2-cyano- <i>p</i> -xylylene	130
2-bromo- <i>p</i> -xylylene	130
dichloro- <i>p</i> -xylylene	130

Once it is in "solution" in the coating, the monomer moves about in random directions by diffusion until it evaporates or is consumed by chemical reaction. The polymer molecules that have already grown to higher molecular weight cannot relocate appreciably owing to entanglement with their neighbors. The rate of diffusion of monomer through the polymer bulk is adequate for the participation of diffusive transport in the mechanism of VDP (ca 10⁻¹⁰ cm² s at room temperature). This can be confirmed in swelling-rate experiments with solvents having similar physical properties, such as

p-xylene.

The monomer is consumed by two chemical reactions: initiation, in which new polymer molecules are generated, and propagation, in which existing polymer molecules are extended to higher molecular weight. In steady-state VDP, both reactions proceed continuously inside polymeric coating, in the reaction zone just behind the growth interface.

The first step of the initiation reaction is the coupling of two monomer molecules to form the dimer diradical (5). The formation of this diradical is energetically uphill, ie, the energy of two benzyl radicals is greater than that of two starting *p*-xylylene systems. The rate of destruction greatly exceeds the rate of formation. Only a trace concentration of the dimer diradical species exists at equilibrium. Further reaction of the dimer diradical with monomer gives more stable diradicals. In these subsequent transformations, a *p*-xylylene is converted into a benzene with a net stabilizing effect. At some stage of oligomerization, the resulting *n*-mer diradical becomes more stable than the *n* *p*-xylylene molecules from which it was constructed. At this point, the new polymer molecule is formed. Thus the overall order of the initiation reaction, the reaction in which new polymer molecules are generated, is some *n* > 2. Initiation chemistry requires no species other than monomer, another unusual aspect of the polymerization chemistry of *p*-xylylene.

The order *n* of the initiation reaction has an important influence on the manner in which the VDP occurs. Because monomer molecules, even in solution at low concentration, are closer together in the condensed phase than they are in the gaseous phase, the rate of initiation is greater in the condensed phase than in the gaseous phase. The higher the order *n*, the more the condensed phase is favored. The order *n*, according to the mathematical model (34) of *p*-xylylene VDP, at the same time governs the effect of monomer pressure on growth rate at a given deposition temperature. The model predicts that growth rate should vary with the pressure raised to the power (*n* + 3)⁴. Thus, if *n* = 3, the growth rate should be proportional to *p*^{1.5}. In an early attempt to determine the pressure dependence of parylene growth rate γ, an expression of γ = *k*·*p*² was reported (35). A pressure exponent of 2 would be interpreted as an initiation order of *n* = 5. Although such a high order would favorably deemphasize "snow," consideration of the energetics of oligomeric *p*-xylylene diradicals would seem to place the order nearer to 3. Perhaps the early investigators did not anticipate a nonintegral order for pressure dependence. A more recent report (36) places *n* at 3 for Parylenes N and C, and 4 for Parylene D. Thus, with *n* > 3, the parylenes are more likely to form a continuous coating than a dust or a snow, the physical form of the product of a gas-phase polymerization. To the extent that snow is included in the formation of a coating (ie, dual-phase polymerization), haze develops.

In the propagation reaction, the monomer molecule reacts with an existing free-radical polymer chain end to make the chain one repeat unit longer. The polymer chains have two active ends, and they grow from both ends at the same time. Under normal coating conditions, the consumption of monomer by propagation must be much higher than its consumption by initiation to obtain high molecular weight polymer. In fact, the number-average molecular weight is determined by the proportion of monomer consumed by the two reactions, and is diminished by increases in deposition temperature or monomer partial pressure.

The concentration of monomer within the coating decreases approximately exponentially with distance from the growth interface. With this decrease in monomer concentration, the rates of initiation and propagation reactions also decrease. Moving back into the polymer from the growth interface, through the reaction zone where polymer is being manufactured, a region in which the polymer formation is essentially complete is gradually entered. Because initiation is of higher order in monomer concentration, it tends to occur closer to the growth interface than does propagation. Under conditions prevailing during a typical deposition, the characteristic depth of the reaction zone is a few hundred nanometers, and the maximum concentration of monomer, ie, the concentration at the growth interface, is of the order of a few tenths percent by weight. Thus the parylene polymerization takes place just behind the growth interface in a medium that is best described as a slightly swollen, solid polymer.

During the vapor deposition process, the polymer chain ends remain truly alive, ceasing to grow only when they are so far from the growth interface that fresh monomer can no longer reach them. No specific termination chemistry is needed, although subsequent to the deposition, reaction with atmospheric oxygen, as well as other chemical conversions that alter the nature of the free-radical chain ends, is clearly supported experimentally.

Polymer Properties. The single most important feature of the parylenes, that feature which dominates the decision for their use in any specific situation, is the vapor deposition polymerization (VDP) process by which they are applied. VDP provides the room temperature coating process and produces the films of uniform thickness, having excellent thickness control, conformality, and purity. The engineering properties of commercial parylenes once they have been formed are given in Table 4. As crystalline polymers, the parylenes retain useful physical integrity up to temperatures approaching their crystalline melting points. However, their glass-transition temperatures, *T*_g, the temperature spans over which the continuous amorphous phase, usually the minority phase, changes from a rigid, vitreous condition to a more flexible, rubbery condition, are probably in the vicinity of ambient temperature. In the case of PPX (Parylene N), careful measurements have established the *T*_g to be centered at 13°C, the range over which *T* affects heat capacity measurements extending from 240 to 330 K (= 33 to +57°C) (24). Because they have similar backbone structures, other parylenes are expected to have similar *T*_g, although no measurements taken with equal care exist. Earlier reports have quoted a somewhat higher (60–90°C) range for the entire family (4). In either case, the parylenes as prepared by their VDP process are further distinguished from conventional polymers in having been generated in a medium that is at least to some extent vitreous.

Table 4. Typical Engineering Properties of Commercial Parylenes

Property	Parylene N	Parylene C	Parylene D	ASTM method
<i>General</i>				
density, g · cm ³	1.110	1.289	1.418	D1505
refractive index, <i>n</i> _D ²³	1.661	1.639	1.669	
<i>Mechanical</i>				
tensile modulus, GPa ^a	2.4	3.2	2.8	D882
tensile strength, MPa ^b	45	70	75	D882
yield strength, MPa ^b	42	55	60	D882
elongation to break, % ^c	30	200	10	D882
yield elongation, % ^c	2.5	2.9	3	D882
Rockwell hardness	R85	R80		D785
coefficient of friction				D1894
static	0.25	0.29	0.35	
dynamic	0.25	0.29	0.31	
<i>Thermal</i>				
melting point, °C	420	290	380	
linear coefficient of expansion at 25°C × 10 ⁵ , K ^{−1}	6.9	3.5		
heat capacity at 25°C, J · (g·K) ^{−1}	1.3 ^d	1.0 ^e		
thermal conductivity at 25°C, W · (m·K) ^{−1}	0.12	0.082		
<i>Electrical</i>				
dielectric constant				D150
60 Hz	2.65	3.15	2.84	
1 kHz	2.65	3.10	2.82	
1 MHz	2.65	2.95	2.80	
dissipation factor				D150
60 Hz	0.0002	0.020	0.004	

1 kHz	0.0002	0.019	0.003	
1 MHz	0.0006	0.013	0.002	
dielectric strength at 25 μm, MV m				D149
short time	275	220	215	
step-by-step	235	185		
volume resistivity at 23°C, 50° RH, Ω	1.4×10^{17}	8.8×10^{16}	2×10^{16}	D257
surface resistivity at 23°C, 50° RH, Ω	1×10^{13}	1×10^{14}	5×10^{16}	D257
	Barrier			
water absorption, %	<0.1	<0.1	<0.1	D570
water vapor transmission at 37°C, ng (Pa·s·m) ^f	0.0012	0.0004	0.0002	E96
gas permeability at 25°C, amol (Pa·s·m) ^g				D1434
N ₂	15.4	2.0	9.0	
O ₂	78.4	14.4	64.0	
CO ₂	429	15.4	26.0	
H ₂ S	1590	26.0	2.9	
SO ₂	3790	22.0	9.53	
Cl ₂	148	0.7	1.1	

^a To convert GPa to psi, multiply by 145,000.
^b To convert MPa to psi, multiply by 145.
^c To convert J to cal, divide by 4.184.
^d Refs. 22 and 27.
^e Ref. 31.
^f To convert ng (Pa·s·m) to g·mil (100 in²·day) at 90° RH, 37°C, multiply by 1240.
^g To convert amol (Pa·s·m) to cm³ (STP)·mil (100 in²·day·atm), multiply by 0.498.

During formation, the motions of the parylene polymer chains in the vitreous medium are restricted. The properties of freshly deposited parylenes, therefore, generally differ from those that have been aged or annealed. Restricted polymer chain motion during VDP severely limits their ability to organize into crystallites, and consequently freshly deposited parylenes are metastable. With the passage of time, and sooner if heated, they will reorganize into a thermodynamically more satisfactory configuration, increasing crystallinity. Certain physical properties of freshly deposited parylenes therefore can be expected to change upon aging or annealing. Of the commercial materials, Parylene C, perhaps as a result of the asymmetry of its repeat unit, is particularly subject to modifications subsequent to its initial formation.

Mechanical Properties. Many of the mechanical properties of the parylenes are similar to those of other conventional plastics. The values in Table 4 are typical of those quoted for the parylenes, but in any particular case can vary with aging or annealing. In an outstanding instance of this effect, the 200° elongation quoted for Parylene C is the value commonly observed on the freshly deposited material. It drops dramatically as crystallinity builds. In general, an increase in crystallinity with aging or annealing results in a lowering of elongation to break and an increase in modulus and strength. The variation in mechanical rigidity of the parylenes as temperature increases is shown in Figure 5, which plots the logarithm of the secant modulus, a measure of stiffness, vs temperature. Generally, a small decrease in rigidity occurs near ambient temperature as the glass transition is traversed. Significant rigidity is then retained up to the point where the crystallites begin to melt.

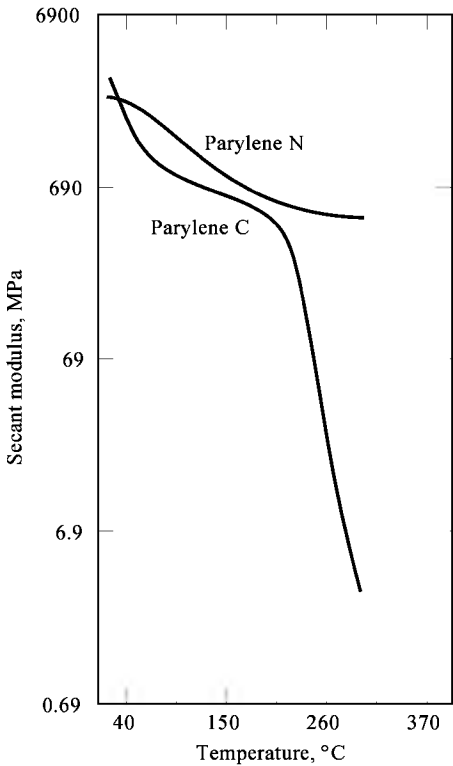


Fig. 5. Stiffness–temperature behavior of Parylenes N and C. To convert MPa to psi, multiply by 145.

It has been reported that Parylene N is deposited in a state of compressive stress (37). The inherent stress is 18 MPa (2300 psi) and is invariant with thickness. This congenital compressive stress can be removed and rendered tensile by a thermal cycle.

Electrical Properties. The bulk electrical properties of the parylenes make them excellent candidates for use in electronic construction. The dielectric constants and dielectric losses are low and unaffected by absorption of atmospheric water. The dielectric strength is quoted for specimens of 25 μm thickness because substantially thicker specimens cannot be prepared by VDP. If the value appears to be high in comparison with other materials, however, it should be noted that the usual thickness for such a measurement is 3.18 mm. Dielectric strength declines with the square root of increasing

thickness. Viewed in this light, the dielectric strength of the parylenes is good, but not outstandingly high. The bulk resistivities are advantageously high because of the purity of the parylenes, their low moisture absorption, and in particular their freedom from the trace ionic impurities present in conventional polymers as residues. Such impurities might be the residues of the catalysts necessary for their formation. The surface resistivities are advantageously high in part because of their low affinity for atmospheric water. The experimental dependence of the electrical properties on temperature is shown in Figure 6.

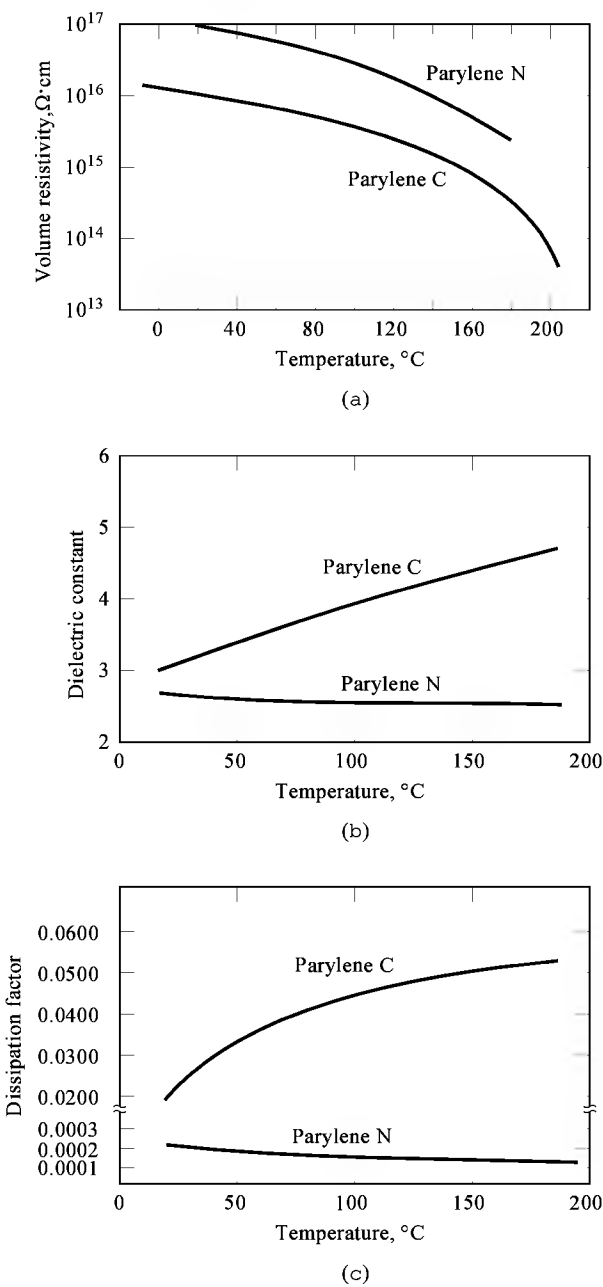


Fig. 6. Variation of electrical properties of Parylenes N and C with temperature.

The question of the value of the dielectric constant of the parylenes in the gigahertz range of frequencies is often of interest to designers of high frequency circuitry. Making such measurements on low loss, low dielectric constant, thin films is experimentally difficult, and no reliable data exist as of this writing (ca 1997). Fortunately, however, an indicator is available. For nonabsorbing, nonmagnetic materials such as the parylenes, the well-known Maxwell relation applies: at any particular frequency, the square of the index of refraction equals the relative permittivity, or dielectric constant (38). The refractive indices for the parylenes at the sodium D line (589 nm), a visible wavelength that corresponds to a frequency of 510 THz ($5.10.167 \times 10^{14}$ Hz), are shown in Table 5, along with their squares and the lower frequency measurement of dielectric constant. Because by virtue of the Maxwell relation the dielectric constants of the parylenes at the much higher frequency of visible light are close to those observed by conventional means, it seems likely that when reliable gigahertz dielectric constant measurements on the parylenes become available, similar values will be obtained.

Table 5. Dielectric Constants and Refractive Indexes of Parylenes

Parameter	Parylene N	Parylene C	Parylene D
dielectric constant			
60 Hz	2.65	3.15	2.84
1 kHz	2.65	3.10	2.84
1 MHz	2.65	2.95	2.80
refractive index			
n_D line	1.661	1.639	1.669
squared	2.76	2.69	2.79

Thermal Properties and Endurance. The heat capacity or specific heat, C_p is a quantity of theoretical thermodynamic significance as well as of practical importance. It has been determined for Parylene N over the temperature range of 220 to 620 K (- 53 to +347 °C) (24,29). Figure 7 gives the results of an experiment in which freestanding films were exposed to constant elevated temperatures in air-circulating ovens for periods of weeks to months; the failure criterion was a 50% loss in tensile strength. Because the test is destructive, each data point (failure time at a given

temperature) required many specimens. In the degradation of many polymers, including the parylenes, tensile strength is maintained until chain scission has reduced molecular weight to the point at which entanglement is no longer a factor in determining physical properties. Beyond that point it drops abruptly. Thus despite the relatively large variance in tensile strength measurements, the 50% loss criterion allows a reasonably precise location of end of useful life on a log time scale (Fig. 8). The data of Figure 7 suggest that Parylenes N, C, and D perform in air without significant loss of physical properties for 10 yr at 60, 80 and 100°C, respectively. Clearly, these data can justifiably be made to serve only as a guide for considering the parylenes for a specific application. Questions of thermal endurance tend to have no clear-cut answers. In situations where performance may be marginal, there is no substitute for retesting under conditions more directly relatable to the application.

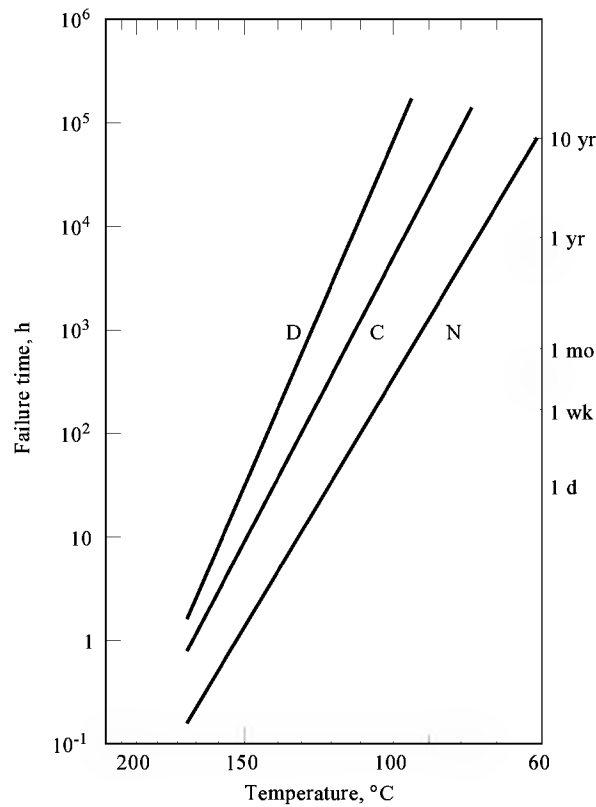


Fig. 7. Useful lifetime of Parylenes N, C, and D as a function of temperature in air. Failure 50% loss in tensile strength.

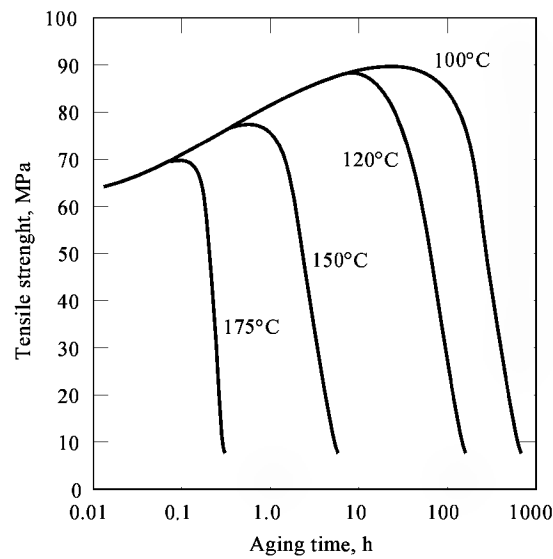


Fig. 8. Tensile strength of Parylene C on aging at elevated temperature. To convert Pa to psi, multiply by 0.145×10^3 .

Degradation. The most important mode of degradation for parylenes is oxidative chain scission. Significantly, hydrolytic degradation is chemically impossible. Oxidative degradation limits the use of parylenes at elevated temperatures in many common applications. Figure 9 shows the effect of exposure to elevated temperature in air or in vacuum on elongation to break, an indicator of toughness or lack of brittleness. The data are given for Parylene C, which suffers substantial change in mechanical properties as the freshly deposited material is annealed. Aging in air at 150°C for ca 100 h causes the elongation to break to drop from its initially very high value to 0, at which point the specimen is mechanically useless. Aging in vacuum at yet higher temperatures (265°C) for a similar length of time gives a stronger, more rigid, stabilized material with 15% elongation at break, the result of annealing.

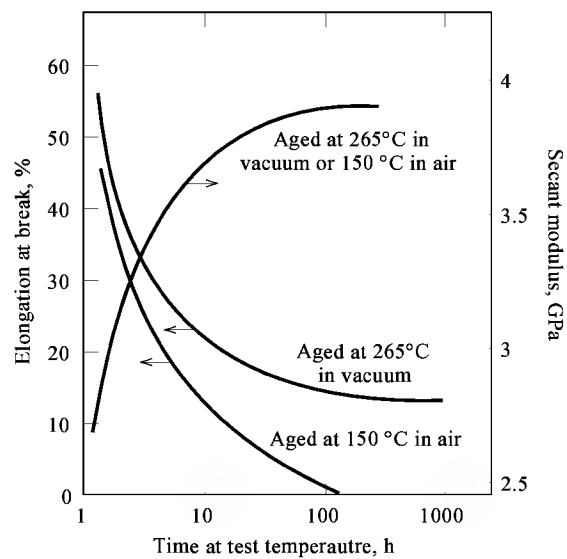


Fig. 9. Effect of temperature on the flexibility of Parylene C in air and vacuum. To convert GPa to psi, multiply by 145,000.

For applications where oxygen can be excluded, eg, in outer space, Figure 10 shows that 10-yr use projections exceed 200°C for Parylenes N, C, and D. Conventional antioxidants, incorporated during or after VDP, can extend the life of the parylenes at elevated temperatures (39).

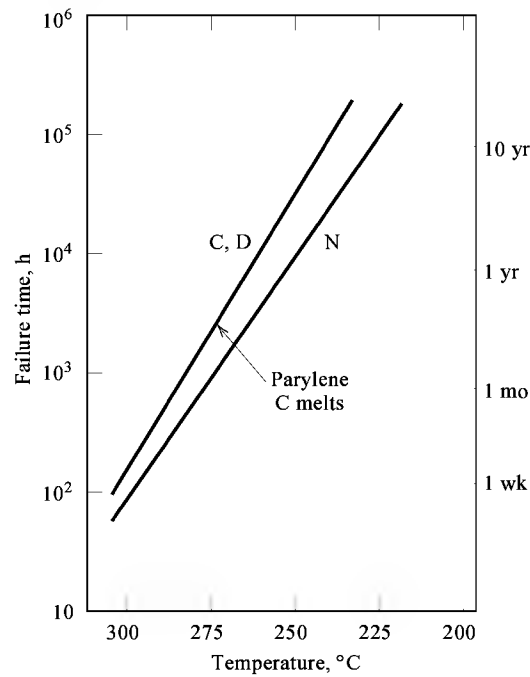


Fig. 10. Long-term effect of aging in vacuum on flexibility of Parylenes C, D, and N at elevated temperature. Failure = 50% loss in tensile strength.

Another factor in oxidative degradation is ultraviolet radiation, of which sunlight is a rich source. The oxidation of parylene appears to be enhanced by ultraviolet radiation. Ozone may play a mechanistic role in the ambient temperature exposure of parylenes to ultraviolet radiation in the presence of oxygen. For the best physical endurance, exposure of the parylenes to ultraviolet light must be minimized.

Barrier Properties. The bulk barrier properties of parylenes are among the best of organic polymeric coatings. The data in Table 4 are the results of tests conducted some time ago, and certain entries seem unaccountably high, particularly the values for the permeability of Parylene N by SO₂ and H₂S. Because film damage might have been the cause of these high test values, experimental confirmation is advisable. More recent values for water transport rates in Parylene C films are available over the temperature range of 20–55°C in a comparative study which includes Mylar A and Kapton H (40). Similar information is provided for Parylene D in a study over the temperature range of 30–80°C (41).

Spectral and Optical Properties. The parylenes do not absorb visible light, and absorb only at the shorter wavelength, high energy end of the ultraviolet range (Fig. 11). Such absorption is expected for the electronic system of the parylenes' benzene ring. Films and coatings are colorless in the visible, becoming opaque to sufficiently short wavelength uv light.

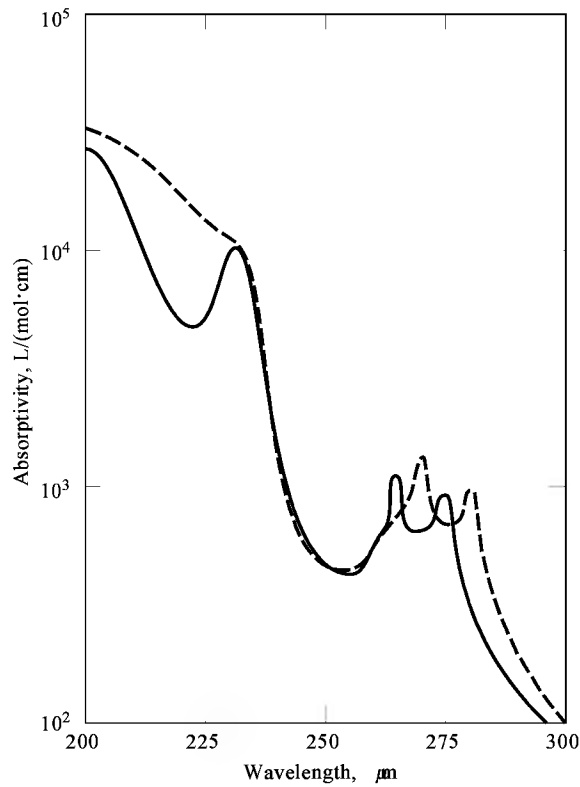


Fig. 11. Ultraviolet absorption spectra of Parylenes N (—) and C (---).

The infrared transmission spectra of Parylenes N, C, and D are compared in Figure 12. Infrared spectra can be used to distinguish among sample films of the three commercial materials should a practical question of identity arise. The spectra in Figure 12 are taken on films of ca 18-μm thickness, which gives a good picture of a typical organic film in the standard infrared presentation format. It is, however, thicker than normally encountered in a typical parylene application. The amount of incident light absorbed at any given wavelength is predictably less than that indicated by the ordinate of Figure 12.

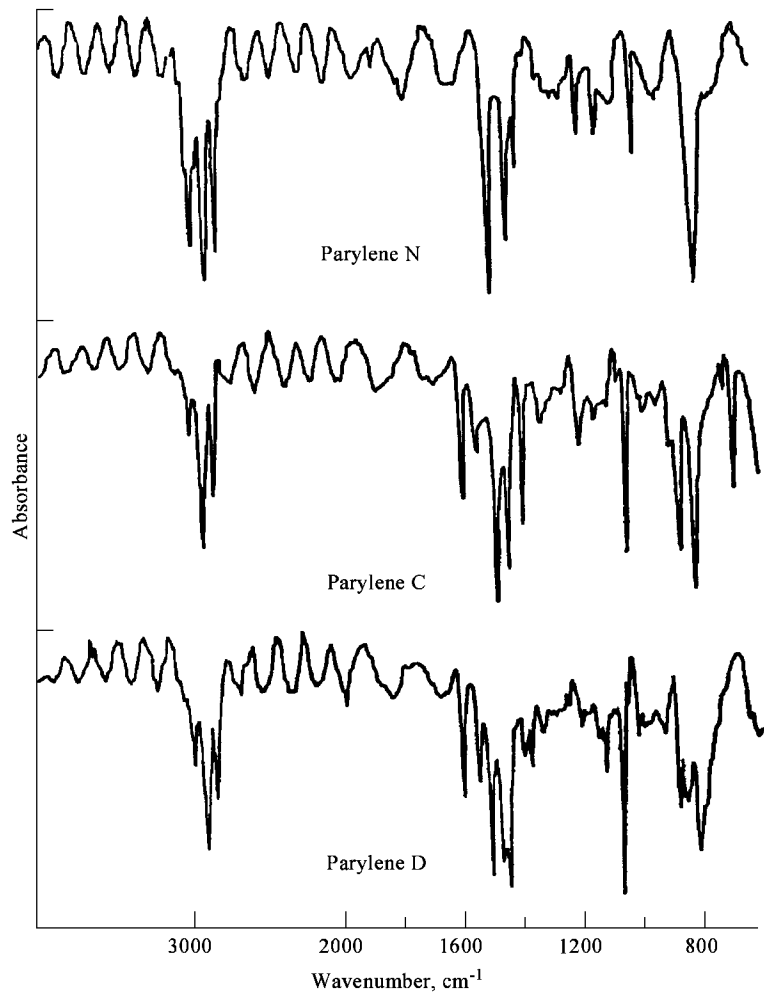


Fig. 12. Infrared absorption spectra of Parylenes N, C, and D films of 18-μm thickness.

Interference effects, which arise because of the extraordinary uniformity of thickness of the film over the spectrometer sample beam, superimposed on the absorption of incident light by parylene films, can be observed. Experimentally, a sinusoidal undulation of the baseline of the spectrum is seen, particularly in the spectral regions where there is little absorption by the sample. These so-called "interference fringe" excursions can amount to some

15–20° transmission and are uniformly spaced in frequency. Larger excursions indicate greater uniformity in thickness. Such fringes are seen toward the left side of the experimental spectra of Figure 12. Although interference phenomena are also encountered experimentally in the visible and ultraviolet regions, their effects have been removed in the consolidation and replotting of the data of Figure 11. In transmission spectra, at the top of the interference fringes (the wavelengths of the sinusoidal maxima), constructive interference occurs in which all incident light that is not absorbed appears for detection. At the fringe troughs (the wavelengths of the sinusoidal minima), a condition of destructive interference exists, at which a portion of the incident beam is reflected back toward the source, thereby escaping detection even though it is not absorbed by the sample. By observing the wavelengths at which constructive interference occurs, the approximate thickness of the sample can be determined by using the expression

$$d = \frac{m \lambda_1 \lambda_2}{2 R (\lambda_1 - \lambda_2)}$$

where λ_1 and λ_2 are any two wavelengths at which interference transmission maxima (ie, constructive fringes) exist, separated by $m-1$ intervening constructive fringes; R is the refractive index of the sample film. The thickness d takes the units of wavelength used for λ . At any condition of constructive interference (fringe), twice the thickness times the refractive index equals an integral number of wavelengths of the incident light. That integer is known as the order of the fringe. Much more precise and accurate thicknesses may be computed from the experimental optical spectra if the order of the fringe is known and if any variation of refractive index with wavelength is taken into consideration.

Unstretched PPX films exhibit an inherent negative birefringence, the optical axis of which is perpendicular to the plane of the film. The refractive index along the optical axis is lower than the refractive index observed in the plane of the film, the difference being $n_e - n_o = 0.075 \pm 0.001$ (42). Where they are observed from the direction normal to the film plane, the films appear to be isotropic. Birefringence is observable only when the films are tilted, or observed in cross section. When stretched, PPX films exhibit a much greater birefringence. The stretch birefringence has an optical axis in the direction of stretch and is positive: $n_z - n_o = +0.2$.

Surface Energy. The surface energies of Parylenes N, C, and D were measured by observing the contact angles for several standard probe liquids. All three have surface energies of approximately 45 mJ / m² (dyn / cm), ie, all test liquids having less than 45 mJ / m² surface tension completely wet the as-deposited parylene surfaces (43).

Plasma treatments using reactive gases (N₂, O₂) as well as inert gases (Ar, He) seemed universally to lower the contact angle for water, an observation that would imply that all such treatments raise the surface energy. Reflectance spectroscopy confirmed the presence of carbonyl in all plasma-treated specimens. Surprisingly, the inert-gas-plasma treatments affected the surface energies of the parylenes more than the reactive gas plasmas did, as indicated by a water contact angle. However, the surface energies of the plasma-treated specimens universally dropped toward the original value on standing in air, but stabilized in about 1 d without recovering the original value. Plasma treatment of parylene surfaces markedly improves the adhesion of polyurethanes, a result that could be in part, at least, the result of the surface energy change.

Crystallinity. The crystallinity of the parylenes determines two of their most important practical characteristics: mechanical strength at elevated temperatures (see Fig. 5) and solvent resistance.

In a manner typical of crystalline polymers, the crystallinity of parylenes is confined to small-submicrometer domains that are randomly dispersed throughout an amorphous continuum. Adjacent polymer chains, in order to acquire greater overall system stability, exhibit a preference to be close to one another, but the extent to which they actually can be is limited by the tangles already present. Because a given polymer chain is long enough to participate successively in several crystallites, these crystallites function as cross-links to strengthen the bulk polymer. This structural role of the crystallites is retained as temperature rises, giving the bulk polymer physical strength even above its glass-transition temperature, where the amorphous phase is changed to a low modulus, rubbery consistency. Because the crystalline domains are much more resistant to permeation than the amorphous phase, they retain their reinforcing structural role even in the presence of permeants in the amorphous phase, thus giving the parylenes their resistance to solvent attack. The crystalline content increases in freshly deposited polymer as it ages, particularly when it is heated.

Parylene N (PPX) possesses a singularly complex crystalline morphology in which two distinct crystalline modifications are recognized. When deposited in the usual VDP fashion, the crystallites tend to be mostly of the α modification. On annealing at a temperature of about 220°C, the α form is converted to the β modification. This transition was originally thought to be irreversible, but studies have demonstrated that it can be made reversible (44). The crystalline phase undergoes further modifications at higher temperatures before reaching its melting point of 420°C; these modifications have not been fully explored. The detailed crystal structures of the α ($a = 592$ pm, $b = 1060$ pm, $c = 655$ pm, $\beta = 134.7^\circ$, for two monomer repeat units per cell) (45) and β ($a = b = 2052$ pm, $c = 655$ pm, $\gamma = 120^\circ$, for 16 monomer repeat units per cell) (46) modifications have been determined.

In Parylene C, the single crystalline form observed is very similar to the α form of Parylene N. Its detailed crystal structure has been determined ($a = 596$ pm, $b = 1269$ pm, c (chain axis) = 666 pm, $\beta = 135.2^\circ$) (47). X-ray studies on the crystal structure of Parylene D have not been reported.

The diffraction pattern produced where x-rays or electrons are directed perpendicularly at the film is the familiar pattern of concentric rings (a powder pattern) produced where the crystallites within the sample are randomly oriented. When the incident beam is directed at an angle to the plane, however, the uniform rings break up into bands or spots, indicative of a preferred orientation of unit cells. Tilted electron diffraction results (48) demonstrate that in Parylene N there is a preference for the $\langle 100 \rangle$ face of the unit cell to lie in the film plane. Because the polymer chain is oriented along the c axis of the unit cell with the benzene rings lying approximately in the plane of the $\langle 100 \rangle$ face, it follows that, at least in among molecules within crystallites, there is a preference for the polymer chains and benzene rings to be oriented in the film plane. Within the film plane, however, the direction of the chains is random. That benzene rings are preferentially oriented in the plane of the film is supported by negative birefringence and by the Raman spectrum of the polymer (49).

Solvent Resistance. At temperatures below the melting of the crystallites, the parylenes resist all attempts to dissolve them. Although the solvents permeate the continuous amorphous phase, they are virtually excluded from the crystalline domains. Consequently, when a parylene film is exposed to a solvent a slight swelling is observed as the solvent invades the amorphous phase. In the thin films commonly encountered, equilibrium is reached fairly quickly, within minutes to hours. The change in thickness is conveniently and precisely measured by an interference technique. As indicated in Table 6, the best solvents, specifically those chemically most like the polymer (eg, aromatics such as xylene), cause a swelling of no more than 3°.

Table 6. Swelling on Immersion in Various Solvents for the Commercial Parylenes at Room Temperature

Solvent	Volume change, °		
	Parylene N	Parylene C	Parylene D
dichlorobenzene	0.2	3.0	1.8
mixed xylenes	1.4	2.3	1.1
monochlorobenzene	1.1	1.5	1.5
2,4-pentanedione	0.6	1.2	1.4
trichloroethylene	0.5	0.8	0.8
acetone	0.3	0.9	0.4
pyridine	0.2	0.5	0.5
isopropyl alcohol	0.3	0.1	0.1
Freon	0.2	0.2	0.2
water, deionized	0.0	0.0	0.0

Applications

Although there is ample evidence in the literature that several industrial groups had a research interest in the PPXs during the 1950s, industrial exploitation was for the most part prevented by the obstacle the parylenes present to conventional fabrication technologies. Because they are generally insoluble in most solvents, even at elevated temperatures, they cannot be used as solvent-based coatings; neither can they be cast as films nor spun as fibers from solution. Because of their high crystalline melting points, melt-working (molding, extrusion, calendering, etc) is also difficult. Yet it is often just these features of solvent resistance and high temperature mechanical strength that constitute the advantages of PPX materials.

It had been recognized from the outset that polymer forms spontaneously on surfaces exposed to the gaseous monomer PX, but the recognition of VDP as an industrially viable process was not immediate. The public announcement, in 1965, of the convenient generation of pure monomer by the Gorham process from the dimer marked the beginning of a period of a gradually increased understanding and acceptance of the unique features of the parylenes.

As a coating technique, VDP offers certain advantages over other coating techniques such as brushing, dipping, or spraying. These advantages stem principally from the fact that the solid coating is formed from the gaseous monomer directly, without an intermediate liquid stage. As a result, the forces of surface tension, which would cause a pulling away from sharp edges and a filling in of troughs in conventional methods, are not operative and therefore do not affect the cross-sectional profile of VDP coatings. Coatings of PPX grow from the substrate surface outward, producing a conformal layer of uniform thickness. Extensive tests (50) demonstrate that the coating so formed is pinhole-free at thicknesses well below those achievable with conventional coating techniques (see COATINGS).

The PPX coatings are formed spontaneously on substrates at or near room temperature; a cure cycle at elevated temperature is not necessary to complete the polymerization chemistry. The substrate need not be subjected to any temperatures above ambient, and no further time is required beyond that needed for the growth of the film. Because the polymerization is spontaneous, no catalyst is necessary. Catalysts that promote other polymerizations are often ionic or ionogenic and, to the extent that they remain in the coating after cure, their residues are capable of participating in charge mobilization, with resulting detrimental effects in electrical properties. Because the coating is formed at room temperature, stresses that might be induced by differential thermal expansion between the temperature of cure and room temperature are avoided. Because PPX coatings are generally much thinner than conventional coatings, any stress that they impart on the substrate is proportionately less.

Circuit Boards. The most important application of parylenes is as a conformal coating for printed wiring assemblies. The parylene VDP process has the ability to produce a continuous, thin, truly conformal coating on geometrically complex, delicate articles. These coatings provide excellent chemical resistance, especially to solvent attack, and resistance to fungal attack. In addition, they exhibit stable dielectric properties over the wide range of temperatures in which military boards are expected to perform, as well as low dielectric constants, which together with thinness minimize the undesirable loading of high frequency tuned circuits. The reliability with which the process operates substantially reduces labor-intensive touch-up and inspection operations otherwise required for conventional coatings.

Parylene C was included among the earliest MIL-I-46058 (51) qualified coatings (as type XY) and has since enjoyed a reputation for superior performance in protecting and preserving the operation of electronic circuits against the detrimental effects of their operating environments. Environmental water is, of course, chief among these adverse factors. Although the rate of water transport through a parylene coating is finite and well documented, the hydrophobic nature of the parylenes makes them excellent barriers to penetration by ionic species. Once the circuit board is clean to the extent that surface ionic contamination is minimized and weak surface layers are eliminated, the parylene coating adheres well to the organic substrate material between the conductors. Ionic conduction along the coating–substrate interface, which would otherwise impair electrical performance in the short run and promote galvanic corrosion and dendritic growths in the long term, is minimized.

The sem photographs of Figure 13 show a conventional FR-4 printed wiring board, coated with an 8-μm film of Parylene C, on which the copper conductor trace has been peeled back. The peeling reveals the rough texture of the epoxy gel coat produced during the original lamination against etched copper in order to achieve maximum laminate peel strength. The region covered with Parylene C in Figure 13 was of the same texture before coating, a fact that, along with parylene's conformality, accounts for much of the texture in the upper surface. Of particular interest is the mode of Parylene C failure as the conductor is peeled. There is evidence of substantial yield prior to failure in the filamentous nature of the failed regions and the web of stretched Parylene C at the apex of the peel. Furthermore, all tensile failure is cohesive, ie, the adhesion of Parylene C to the underlying epoxy gel coat is greater than the cohesive strength within the coating material. The higher magnification (Fig. 13b) shows the line of demarcation between the regions covered with copper and with Parylene C prior to the peel, revealing that the breadth of the failed Parylene C exceeded by far its original thickness.

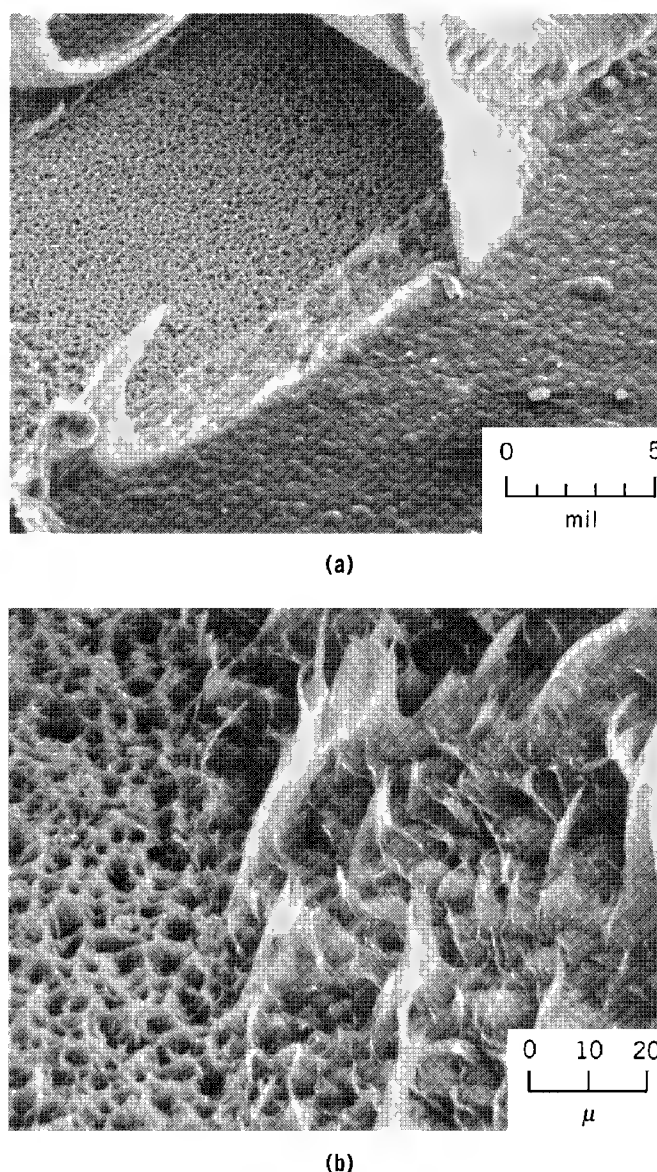


Fig. 13. Scanning electron microscope (sem) photographs of Parylene C-coated printed circuit conductor peeled to demonstrate the adhesion of the coating to the substrate. To convert mil to meters, multiply by 2.54×10^{-5} .

The circuit board with all its components attached presents a variety of surface types for the conformal coating, ranging from plastics through ceramics and glass to metals, in particular aluminum and tin-lead solder. No single mechanism for adhesion can serve all sites effectively. Of particular importance for the epoxy board are the surfaces between conductor traces. Because epoxy is an organic material that is permeable to PX monomer, the growth mechanism of the parylene coating provides adhesion by the interpenetrating network it produces. Care must be taken in cleaning the board surface before coating of leakage-producing ionic contamination as well as organic soils, which interfere with a secure xylylene interpenetration lock. In the cases of ceramic and metal surfaces, monomer penetration is impossible, and other adhesion mechanisms must be engaged. Adhesion to metals is often not essential to the performance of the board, but it is usually desirable. Adhesion to ceramics or glasses is vital in situations where electrical leakage across their surfaces is to be minimized, and is particularly necessary in hybrids. Because the outside surface of common metals is an adherent coating of native oxide, adhesion to metals, ceramics, and glasses can often be achieved using the same approach, ie, treatment with a compatibilizing layer of an organosilane (52). γ -Methacryloxypropyltrimethoxysilane (A-174) is the organosilane most commonly used for the purpose. In action, the trialkoxysilane end of the molecule hydrolyzes and bonds covalently with the oxide on the substrate surface. The methacrylic end of the molecule provides a hydrophobic surface to accommodate parylene deposition, as well as the opportunity to form covalent bonds with the *p*-xylylene polymer through its double bond. It is, of course, essential to provide no more than a monolayer of the organosilane. Water plays an important role in the bonding chemistry between the organosilane and the oxide, and the treatment is often applied from aqueous solution. An alternative treatment is by exposure of the substrate circuit boards to pure gaseous organosilane (53), a process which often takes several days to develop optimum adhesion.

Fluorescence is frequently required in a circuit-board coating to assist inspection of the board after it is coated. Fluorescent parylene coatings can be prepared by admixing a fluorescent agent into the dimer charge. The agent must have just the right volatility to pass through the process and deposit with the polymer, yet it must be sufficiently stable that it survives the conditions it encounters during the process. Agents that have been used in this manner are anthracene and certain fluorescent whiteners of the Calcofluor family, the 7-dialkylamino-4-methylcoumarins (see FLUORESCENT WHITENING AGENTS).

Hybrid Circuits. The use of parylenes as a hybrid circuit coating is based on much the same rationale as its use in circuit boards. A significant distinction lies in obtaining adhesion to the ceramic substrate material, the success of which determines the eventual performance of the coated part. Adhesion to the ceramic must be achieved using adhesion promoters, such as the organosilanes.

In the coating of hybrid circuits, parylene provides certain other advantages. In certain technologies, chips are mounted with substrate clearances of as little as 10 μm . The parylene VDP process penetrates this narrow space, coating the underside of the chip as well as the substrate opposite it. Conventional coatings cannot flow into such a narrow space and must rely on simply sealing it off, leaving an air pocket under the chip. Where the fabrication technology uses wire bonds to connect the chips with substrate conductor patterns electrically, the fine (25- μm dia) wires are coated by the parylene VDP process to the same thickness as all other surfaces. A marked increase in wire-bond strength in pull tests on coated hybrids is observed. The increase in strength can be attributed to the strengthening effect of the coating on the wire (which typically exceeds in cross section the amount of metal in the wire) or to a reinforcement of the welded junction to the chip pad or substrate conductor.

A hybrid solid-state relay needed for a NASA space experiment underwent redesign to provide 2500-V d-c isolation, input-to-output, while undergoing temperature cycling from -120 to $+180^\circ\text{C}$, as detailed in an instructive published case history (54). The electrical isolation requirement was met only by using a conformal coating, although several conventional coatings were evaluated along with parylene. These stresses in the thicker conventional coatings associated with temperature excursions resulted, in all cases other than parylene, in the rupture of the circuitry and the functional failure of the

relay.

During the manufacture of hybrid circuits, it is possible to generate small metallic fragments such as wire chips and solder balls. If they are not removed before hermetic packaging, they can disrupt performance of the unit at some future time. In certain aerospace or military applications where no physical replacement of the unit is possible, such a failure could mean the loss of a mission. A program intended to improve the reliability of hermetic packages in critical systems concluded by finding a solution to this problem by applying a parylene coating to the inside of the hybrids. The coating confers electrical insulation to all surfaces and tends to anchor any loose particles; it also confers nonconductance should they break loose from their moorings. From the standpoint of coating technology, the most interesting part of this solution to the particle-immobilization problem is the throwing power, demonstrated by the fact that the entire internal surface of the hybrid is coated through a 0.5–1.0-mm hole in the lid of the hermetic package (55).

Semiconductors. The distinctive conformality of parylene is often regarded as a handicap for meeting the requirements of an interlayer dielectric (ILD) in large-scale integration (LSI) multilevel interconnection systems. After many layers of successive patterning, a planarizing procedure is very much needed, not a procedure that would replicate the existing lumpiness. A polyimide, the first organic compound to be used in a commercial semiconductor structure, was selected for this application (56), because of the planarization inherent in its spinning application procedure. However, because parylene offers distinct advantages in purity and thickness control and low moisture absorption (particularly in comparison to the polyimides), it receives continued attention in the search for new fabrication techniques (57). Recognition of the need to lower the dielectric constant of ILD insulating materials as a means of improving device performance has generated a renewed interest in organics in general, and the parylenes in particular. To meet this opportunity, the intent to produce the dimer of the per- α -fluorinated version known as Parylene AF4 (poly($\alpha,\alpha,\alpha',\alpha'$ -tetrafluoro-*p*-xylylene)) [74952-03-7] has been announced (58). In addition to offering a dielectric constant in the vicinity of 2.35, its thermal stability is such that it can withstand back-end-of-line processing steps at temperatures up to 450°C without perceptible weight loss.

In an interesting extension of the conventional solid-state device concept, where a thin film covers the channel of a field-effect transistor (FET), the electron current in the channel may be made to respond to changes in the chemical environment on the opposite side of the film. Thus a silicon nitride film on a channel gives an ion-selective FET (ISFET) that responds to the pH of a solution with which it is in contact. The hydrophobic nature of parylene, on the other hand, enables the channel it covers to be independent of solution composition, and such a channel can be used as a reference electrode. An important advantage of parylene in this context is the thinness with which it can be deposited as a continuous film. The gate coatings are 0.1 μm thick, a thinness required if the sensor is to have a speedy response. An experimental probe-type all-solid-state pH sensor has been demonstrated (59). It was also demonstrated that chemically modifying the parylene surface with crown ether compounds can give an ISFET (known as a CHEMFET) that responds to potassium ion concentration.

Sensors for physical effects are also fabricated using standard semiconductor technology. A notable example is a silicon membrane pressure transducer, which, among its many other uses, is currently employed as a manifold pressure transducer in automobiles (60). The thin membrane of single-crystal silicon was prepared by etching most of the wafer thickness away, into which a resistive Wheatstone Bridge network that enables an electrical readout of the membrane's physical strain while flexing had previously been diffused. Although it is desirable to insulate the membrane electrically from the medium it measures, it is also desirable to do so without affecting the elasticity of the membrane, so that it retains its calibration of electrical output vs input pressure. A thin coating of Parylene C easily accomplishes both. In an inversion of roles, these same pressure transducers were used to sense the stresses imparted to coated substrates by a variety of MIL-I-46058C qualified coating materials (61). Whereas all other coatings, as a result of their cure shrinkages acting through their greater thicknesses modified by their assorted elastic moduli, affected sensor calibration in all cases, the presence of 13 μm of Parylene C was barely detectable.

Capacitors. The outstandingly low dielectric loss of parylenes make them superior candidates for dielectrics in high quality capacitors. Furthermore, their dielectric constant and loss remain constant over a wide temperature range. In addition, they can be easily formed as thin, pinhole-free films. Kemet Flatkaps are fabricated by coating thin aluminum foil with Parylene N on both sides and winding the coated foils in pairs (62).

Parylenes are also used to coat the rotor and stator plates of miniature variable air-gap tuning capacitors (63). Coating the plate raises the voltage that must be applied between two parallel plates for a discharge to occur in the air gap. Thus, a closer plate spacing is permitted, and a smaller unit having the same electrical function can be built. This is also beneficial in improving design capacitance per unit volume. Compared to the alternative of interleaving dielectric films between the plates, the coated-plate technique offers the advantages of simplicity in assembly as well as reliability of the unit. Uniformity and thickness control are of paramount importance in this case.

The Bathythermograph. The thermistor sensing probe of a disposable bathythermograph is coated with parylene. This instrument is used to chart the ocean water temperature as a function of depth. Parylene provides the needed insulation resistance and is thin and uniform enough to permit a rapid and accurate response to the temperature of the surrounding salt water (64).

Miniature Electrical Components. In the manufacture of miniature transformers and motor armatures, it is often necessary that the winding be electrically isolated from the bobbin. The insulating varnish of fine copper wire often does not survive the rigors of the winding operation, and, particularly in the case of the smaller devices, the bobbin requires insulation. However, a thicker insulation than is absolutely necessary takes up space that otherwise could be used for more turns of wire. Thus, thickness control in the coating of the bobbin means better performance of the finished device. Parylene is used in the manufacture of high quality miniature stepping motors, such as those used in wristwatches, and as a coating for the ferrite cores of pulse transformers, magnetic tape-recording heads, and miniature inductors, where the abrasiveness of the ferrite is particularly damaging. In the coating of complex, tiny objects such as these, the VDP process has an extra labor-saving advantage. It is possible to coat thousands of such articles simultaneously by tumbling them during the VDP operation (65).

Medical Uses. Under the auspices of the National Institutes of Health Artificial Heart program (66), Parylene C received considerable attention as a component of a novel scheme for achieving blood compatibility in a flexible pump chamber wall by anchoring the endothelial tissue of the host to it. A microfiber nonwoven fabric anchored to the flexible wall by an overcoating of Parylene C provides scaffolding for cellular ingrowth (67). Although the scheme is not used today, the program contributed a significant amount of favorable biocompatibility data on Parylene C, particularly in the area of cytotoxicity (68). Parylene's minimal perturbation of cells growing in its vicinity can be ascribed to its high purity and its ability to slow down impurity species that might otherwise diffuse out of a substrate material. The corrosive biological environment does not affect parylene, which cannot be hydrolytically degraded.

Parylene's use in the medical field is linked to electronics. Certain pacemaker manufacturers use it as a protective conformal coating on pacemaker circuitry (69). The coated circuitry is sealed in a metal can, so that the parylene coating serves only as a backup should the primary barrier leak. There is also interest in its use as an electrode insulation in the fabrication of miniature electrodes for long-term implantation to record or to stimulate neurons in the central or peripheral nervous system, as the "front end" of experimental neural prostheses (70). One report describes the 3-yr survival of functioning parylene-coated electrodes in the brain of a monkey (71).

Artifact Conservation. As books age, the paper of their pages becomes brittle. A relatively thin coating of parylene can make these embrittled pages stronger (72). Parylene coats the fibers conformally and welds them together at crossing points, providing added structural strength. Although parylene does little to retard the chemical processes that make the paper brittle, it restores some physical strength. Furthermore, the parylene coating can be applied to the existing book without disturbing the binding, thus saving the labor of disassembly. The concept has been extended to other fragile artifacts, such as fabrics.

Laser Fusion Targets. In the search for new energy sources, some research is directed toward the thermonuclear fusion of deuterium–tritium (DT) mixtures. In laser-driven inertial confinement fusion, a single laser pulse crushes a target containing the DT fuel to one thousandth of its normal volume and achieves a temperature of 10^8 K. The energy per pulse in current lasers is small, which limits the experimental targets to a diameter of about 100 μm . The outer layer of the spherical target absorbs the omnidirectionally incident laser energy and ablates as it is thermally destroyed, imparting a reactive compressive force on the inner portions of the target, thus compressing it. The dimensions of the outer layer are central to achieving hydrodynamic stability during the implosion process. Concentricity, sphericity, and thickness (ca 10 μm) must be better than 5%, and surface roughness must be no greater than 0.1 μm . Parylene is a leading contender for the outer layer, and considerable ingenuity has been applied in several methods for experimental target fabrication (73).

Contamination Control. In a number of developing technologies, contamination by small particles is a serious problem. To the extent that

the generation of freely mobile particles is reduced by securing them to their initial locations, parylene coating of critical system components can be useful. In the manufacture of Winchester disk drive units, large magnesium castings are coated with parylene. This increases system reliability (74), presumably because the large surface within the sealed unit, if not coated, is capable of producing particles of the same destructive potential that the system seal is supposed to prevent from entering.

Barrier Coating. The bulk permeabilities of parylenes, although finite, are generally lower than those of most other types of plastic materials. The further advantage of coating continuity inherent in the VDP process allows the parylenes to realize the benefits of their low permeabilities to the fullest, without leakage of the permeant through coating defects such as cracks or pinholes. Thus the parylenes are uniquely suited as protective encapsulants or barrier coatings (see BARRIER POLYMERS).

Where pieces of lithium metal are coated with Parylene C, and their subsequent reaction to water vapor follows, the steady-state generation of hydrogen can be controlled exactly by the rate of water transport through the coating (75). Pellets of nitronium perchlorate, a potent oxidizer useful as a solid-rocket propellant component, can be rendered less moisture-sensitive and compatible with organic binders by applying a coating of Parylene C (76). The absorption of water by particulate ammonium nitrate could be reduced tenfold by as little as a 0.7% coating of Parylene C (77). Particles of ammonium nitrate remain free-flowing after long exposures to ambient conditions of temperature and moisture when coated with as little as 0.2% Parylene C. The thermal sensitivity of Parylene C-coated ammonium nitrate (time to explosion) is unaffected by the coating. On the other hand, the 1.5–8% coating of particulate military-grade RDX explosive, cyclotrimethylene trinitramine, produced significant changes in physical and explosive properties. The changes were attributed to the chemical properties of the protective Parylene C coating and to the virtual absence of encapsulated agglomerates of RDX crystals (78).

In a recent report (79), a 150–200 mg/cm² Parylene C coating provided protection against moisture uptake by three-phase, polyimide, microballoons, and air, syntactic foams. In a previously reported coating of a similar foam, the stated purpose was strengthening (80).

Corrosion Control. The oxygen and water permeability and thinness of the parylene coatings notwithstanding, the rate of corrosion of a coated surface is often significantly reduced. In one case, the corrosion of a plated wire memory was reduced to the point where no bits were lost during 30 cycles of a MIL STD 202D-106C test by overcoating the permalloy plating, which had been previously coated with a Ni–P anticorrosive layer, with Parylene N (81).

Dry Lubricant. The static and dynamic coefficients of friction for the parylenes are low and virtually the same. This feature is an advantage in the use of a parylene coating as a dry lubricant on the bearing surfaces of miniature stepping motors. Coating a threaded ferrite core significantly reduces the abrasion to coil forms (82).

Pellicles and Membranes. By separating the coating from the substrate after deposition, the unique coating features of parylenes, especially continuity and thickness control and uniformity, can be imparted to a freestanding film. In practice, a sheet of smooth glass is coated with a layer of a hygroscopic substance before being coated with parylene. The film is then lifted from the glass by water immersion to activate the release agent. In this manner, uniform, continuous, freestanding parylene films with thicknesses of less than 0.1 μm can be prepared. Applications of such films include optical beam splitters (83), a window for a micrometeoroid detector (84), a detector cathode for an x-ray streak camera (85), windows for x-ray proportional counters (86), a charge stripper for converting a portion of the H-output beam of a 50 MV LINAC to neutral H⁰ for diagnostic purposes (87), and a massless support for projectile abrasion testing (88); proposals have included the membrane structure for a solar sail (89).

Health and Safety

In a world increasingly conscious of the dangers of contact with chemicals, a process that is conducted within the walls of a vacuum chamber, such as the VDP process for parylene coatings, offers great advantages. Provided the vacuum pump exhaust is appropriately vented and suitable caution is observed in cleaning out the cold trap (trace products of the pyrolysis, which may possibly be dangerous, would collect here), the VDP parylene process has an inherently low potential for operator contact with hazardous chemicals.

To an experienced operator trained in the handling of industrial chemicals, the dimers present little cause for concern in handling or storage. The finished polymer coating presents even less of a health problem; contact with the reactive monomer is unlikely. In the ancillary operations, such as cleaning or adhesion promotion, the operator must observe suitable precautions. Before using the process chemicals, operators must read and understand the current Material Safety Data Sheets, which are available from the manufacturers.

BIBLIOGRAPHY

"Xylylene Polymers" in *ECT* 3rd ed., Vol. 24, pp. 744–771, by S. M. Lee, Ford Aerospace and Communications Group.

1. L. A. Errede and M. Szwarc, *Q. Rev. Chem. Soc.* **12**, 301 (1958); M. Szwarc, *Polym. Eng. Sci.* **16**(7), 473 (1976); W. F. Gorham and W. D. Niegisch "Xylylene Polymers" in N. M. Bikales, ed., *Encyclopedia of Polymer Science and Technology*, Vol. 15, Interscience Publishers, a Division of John Wiley & Sons, Inc., New York, 1989, pp. 98–124.
2. **U.S. Pat. 3,342, 754** (Sept. 19, 1967), W. F. Gorham (to Union Carbide Corp.).
3. W. F. Gorham, *J. Polym. Sci. A-1* **4**, 3030 (1966).
4. W. F. Gorham, *J. Polym. Sci. A-1* **4**, 3027 (1966).
5. R. B. Woodward and R. Hoffmann, *The Conservation of Orbital Symmetry*, Chapt. 6, pp. 70, 85, and reference 96 therein.
6. L. K. Montgomery and co-workers, *J. Am. Chem. Soc.* **108**(19), 6004 (1986).
7. P. G. Mahaffy, J. D. Wieser, and L. K. Montgomery, *J. Am. Chem. Soc.* **99**(13), 4514 (1977).
8. **U.S. Pat. 4,532,369** (July 30, 1985), H. Hartner (to Merck); **U.S. Pat. 4,769,505** (Sept. 6, 1988), C. Lee and D. R. Bassett (to Union Carbide Corp.).
9. D. J. Williams, J. M. Pearson, and M. Levy, *J. Am. Chem. Soc.* **92**, 1436 (1970).
10. S. K. Pollak, B. C. Raine, and W. J. Hehre, *J. Am. Chem. Soc.* **103**(21), 6308 (1981).
11. M. J. S. Dewar, *J. Am. Chem. Soc.* **104**(5), 1449 (1982).
12. H. G. Gilch, *J. Polym. Sci. A-1* **4**, 438–439 (1966).
13. V. V. Korshak and S. L. Sosin, *Vysokomol. Soyed.* **7**(2), 232–238 (1965).
14. **U.S. Pat. 3,787,382** (Jan. 22, 1974), A. N. Wright, W. R. Burgess, and E. V. Wilkus (to General Electric Co.); M. Inoue, H. Fujioka, T. Sorita, and T. Tanaka, *ACS Polym. Preprints* **28**(2), 332 (1987).
15. Y. Takai, Y. Hayase, T. Mizutani, and M. Ieda, *J. Phys. D.* **17**, 399 (1984).
16. J. R. Salem, F. O. Sequeda, J. Duran, W. Y. Lee, and R. M. Yang, *J. Vac. Sci. Technol.* **A4**(3), 369 (1986); Y. Takahashi, M. Iijima, K. Inagawa, and A. Itoh, *ibid.* **A5**(4), 2253 (1987).
17. F. E. Arnold and L. S. Tan, *31st Int. SAMPE Symp.* **31**, 1185 (1986); **U.S. Pat. 4,711,964** (Dec. 8, 1987), L. S. Tan and F. E. Arnold (to University of Dayton).
18. Y. Ito, S. Miyata, M. Nakatsuka, and T. Saegusa, *J. Org. Chem.* **46**, 1043 (1981).
19. **Eur. Pat. 220,743** (May 6, 1987), R. Ungarelli, M. A. Baretta, and L. Sogli (to Montedison SpA).
20. **Eur. Pat. 226,255** (June 24, 1987), R. Ungarelli, M. A. Baretta, and A. Malacrida (to Montedison SpA).
21. **Eur. Pat. 220,744** (May 6, 1987), G. F. Pregaglia, M. A. Baretta, and A. Malacrida (to Montedison SpA).
22. C. J. Brown, *J. Chem. Soc.*, 3265 (1953).
23. R. H. Boyd, *Tetrahedron* **22**, 119 (1966); D. L. Rodgers, E. F. Westrum, Jr. and J. T. S. Andrews, *J. Chem. Thermodyn.* **5**, 733–739 (1973).
24. D. E. Kirkpatrick, Ph.D. dissertation, Rensselaer Polytechnic Institute, Troy, N.Y., 1984.

25. M. Gazicki, G. Surendran, W. James, and H. Yasuda, *J. Polym. Sci. Polym. Chem. Ed.* **23**, 2255 (1985).

26. H. Sawada, *Thermodynamics of Polymerization*, Marcel Dekker, Inc., New York, 1976.

27. M. Souders and co-workers, *Ind. Eng. Chem.* **41**, 1048 (1949).

28. D. R. Stull, E. F. Westrum, Jr., and G. C. Sinke, *The Chemical Thermodynamics of Organic Compounds*, Krieger, Malabar, Fla., 1987, p. 699.

29. D. E. Kirkpatrick, L. Judovits, and B. Wunderlich, *J. Polym. Sci. Polym. Phys. Ed.* **24**, 45 (1986).

30. R. W. Warfield and M. C. Petree, *J. Polym. Sci.* **55**, 497 (1961).

31. S. Y. Lazareva, *Vysokomol. Soedin.* **A21**(7), 1509 (1979).

32. B. L. Joesten, *J. Appl. Polym. Sci.* **18**, 439 (1974).

33. W. A. Alpaugh and D. R. Morrow, *Thermochim. Acta* **9**, 171 (1974).

34. W. F. Beach, *Macromolecules* **11**(1), 72 (1978).

35. F. E. Cariou, D. J. Vally, and W. E. Loeb, *IEEE Trans. Parts Mater. Packag.* **PMP-1**(1), 54 (1965).

36. J. F. Gaynor, S. B. Desu, and J. J. Senkevich, *Macromolecules* **28**, 7343–7348 (1996).

37. B. J. Bachman, *Proceedings of the 1st International SAMPE Electronics Conference*, Santa Clara, Calif., June 23–25, 1987, pp. 431–440.

38. A. von Hippel in E. U. Condon and H. Odishaw, eds., *Handbook of Physics*, 2nd ed., McGraw-Hill Book Co., Inc., New York, 1968, Pt. 4, p. 104.

39. **U.S. Pat. 4,163,828** (Aug. 7, 1979); **U.S. Pat. 4,173,664** (Nov. 6, 1979), G. S. Cieloszyk (to Union Carbide Corp.); **U.S. Pat. 4,176,209** (Nov. 27, 1979), T. E. Baker, G. L. Fix, and J. S. Judge (to Raytheon); T. E. Nowlin, D. F. Smith, and G. S. Cieloszyk, *J. Polym. Sci. Polym. Chem. Ed.* **18**(7), 2103 (1980); T. E. Baker, G. L. Fix, and J. S. Judge, *J. Electrochem. Soc.* **127**(8), 1851 (1980).

40. W. H. Hubbel, Jr., H. Brandt, and Z. A. Munir, *J. Polym. Sci. Polym. Phys. Ed.* **13**, 493 (1975).

41. E. Sacher, *J. Appl. Polym. Sci.* **28**, 1535 (1983).

42. R. S. Corley, H. C. Haas, M. W. Kane, and D. I. Livingston, *J. Polym. Sci.* **13**, 137 (1954).

43. T. E. Nowlin and D. F. Smith, *J. Appl. Polym. Sci.* **25**, 1619 (1980).

44. D. E. Kirkpatrick and B. Wunderlich, *J. Polym. Sci. Polym. Phys. Ed.* **24**, 931 (1986).

45. R. Iwamoto and B. Wunderlich, *J. Polym. Sci. Polym. Phys. Ed.* **11**, 2403 (1973).

46. S. Isoda, M. Tsuji, M. Ohara, A. Kawaguchi, and K. Katayama, *Polymer* **24**, 1155 (1983).

47. S. Isoda, T. Ichida, A. Kawaguchi, and K. Katayama, *Bull. Inst. Chem. Res. Kyoto Univ.* **61**(3), 222 (1983).

48. W. D. Niegisch, *J. Appl. Phys.* **38**(11), 4110 (1967).

49. M. S. Mathur and G. C. Tabisz, *J. Cryst. Mol. Struct.* **4**, 23 (1973).

50. M. A. Spivack and G. Ferrante, *J. Electrochem. Soc.* **119**(11), 1592 (1969).

51. *MIL-I-46058C(6): Insulating Compound, Electrical (for Coating Printed Circuit-Assemblies)*, U.S. Dept. of Defense, Washington, D.C., Nov. 8, 1982.

52. **U.S. Pat. 3,600,216** (Aug. 17, 1971), D. D. Stewart (to Union Carbide Corp.).

53. **Ger. Pat. 2,737,792** (Mar. 2, 1978), D. M. Mahoney and W. F. Beach (to Union Carbide Corp.).

54. B. L. Slater, T. J. Riley, and W. Janssen, *Proceedings of the IEEE Power Electronics Specialists Conference*, Pasadena, Calif., June 11–13, 1973, pp. 163–169.

55. V. S. Kale and T. J. Riley, *IEEE Trans. Parts Hybrids Packag.* **PHIP-1**(3), 273 (1977).

56. D. L. Bergeron, J. P. Kent, and K. E. Morrett, *22nd Proceedings of the International Reliability Physics Symposium*, Las Vegas, Nev., 1984, pp. 229–233.

57. J. M. Baker, J. H. Magerlein, and M. J. Palmer, *IBM Tech. Discl. Bull.* **27**(7B), 4382 (Dec. 1984); **Jpn. Pat. 62 106456** (Nov. 1, 1986), M. Kanazawa and S. Kawanako (to Fujitsu, Ltd.); W. F. Beach and T. M. Austin, *Hybrid Circ. Technol.*, 33–36 (Jan. 1990).

58. J. Wary, R. A. Olson, and W. F. Beach, *2nd International Dielectrics for VLSI/ULSI Multilevel Insulation Conference (DUMIC)*, Santa Clara, Cal., Feb. 20–21, 1996, pp. 207–213.

59. T. Matsuo, H. Nakajima, T. Osa, and J. Anzai, *Sensors Actuat.* **9**(2), 115 (1986).

60. R. Olson, *Electron. Packag. Prod.*, 213 (May 1980).

61. R. Olson, *Circuits Mfg.*, 57 (Jan. 1986).

62. **U.S. Pat. 3,327,184** (June 20, 1967), D. J. Valley (to Union Carbide Corp.); **U.S. Pat. 3,319,141** (May 9, 1967), F. E. Cariou, W. E. Loeb, and D. J. Valley (to Union Carbide Corp.).

63. **U.S. Pat. 3,949,280** (Apr. 6, 1976), S. Odagiri, M. Nuka, N. Shinba, M. Baba, and Y. Iizuka (to Mitsumi Electric Co., Ltd.); **Ger. Pat. 2,546,332** (Apr. 29, 1976), J. Lefeber and J. deJonge (to N. V. Philips Gloeilampenfabriken).

64. **U.S. Pat. 3,389,604** (June 25, 1968), G. B. Williams (to Buzzards Corp.).

65. **Ger. Pat. 2,329,186** (Jan. 3, 1974), F. R. Tittmann (to Union Carbide Corp.); **U.S. Pat. 3,300,332** (Jan. 27, 1967), W. F. Gorham and H. L. Willard (to Union Carbide Corp.).

66. *Contract No. N01-HV-8-1388*, National Heart, Lung and Blood Institute, Division of Heart and Vascular Diseases, Bethesda, Md.

67. F. R. Tittmann and W. F. Beach in M. Szycher and W. J. Robinson, eds., *Synthetic Biomedical Polymers: Concepts and Applications*, Technomic Publishing Co., Inc., Westport, Conn., 1980, pp. 117–132.

68. R. H. Kahn and W. E. Burkel, *In Vitro* **8**(6), 451 (1973).

69. R. Olson, *Electron. Packag. Prod.*, 213 (May 1980).

70. E. M. Stemmed, *J. Electrophysiol. Tech. Eng.* **10**(1), 19 (1983); G. E. Loeb, M. Salcman, and E. M. Schmidt, *IEEE Trans. Biomed. Eng.* **BME-24**(2), 121 (1977).

71. E. M. Schmidt, J. S. McIntosh, and M. J. Bak, *Med. Biolog. Eng. Comput.*, 96 (Jan. 1988).

72. B. J. Humphrey, *J. Am. Inst. Conserv.* **25**(6), 15 (1986); C. J. Shahani and W. K. Wilson, *Am. Sci.* **75**(3), 240 (1987).

73. D. F. Peiffer, T. J. Corley, G. M. Halpern, and B. A. Brinker, *Polymer* **22**(4), 450 (1981); R. Q. Gram, H. Kim, J. F. Mason, and M. Wittman, *J. Vac. Sci. Technol.* **A4**(3), 1145 (1986); K. W. Beig, *J. Vac. Sci. Technol.* **18**(3), 1231 (1981); **U.S. Pat. 4,381,963** (May 3, 1983), I. S. Goldstein, F. D. Kalk, and H. W. Deckman (to The University of Rochester).

74. R. Olson and R. Veague, *Microcontamination*, 57 (Jan. 1985).

75. M. A. Spivack, *Corrosion (Houston)* **26**(9), 371 (1970).

76. **U.S. Pat. 3,523,839** (Aug. 11, 1970), L. Shechter and W. E. Loeb (to Union Carbide Corp.); **U.S. Pat. 3,556,881** (Jan. 19, 1971), W. F. Gorham and W. E. Loeb (to Union Carbide Corp.).

77. T. C. Castorina and A. F. Smetana, *J. Appl. Polym. Sci.* **18**, 1373 (1974).

78. A. F. Smetana and T. C. Castorina, *Proceedings of the Propellants, Explosives and Pyrotechnics Conference*, Dec. 3–4, 1974, Picatinny Arsenal, Dover, N.J.

79. A. Calahorra, O. Gara, and S. Kenig, *J. Cell. Plast.* **23**(7), 383 (1987).

80. R. J. McWhirter, *Report from Bendix Corp.*, BDX-613-2358, Bendix Corp., Kansas City, Mo., Sept. 1980.

81. K. Makino, S. Kawakami, and S. Orihara, *Fujitsu Sci. Tech. J.* **9**(4), 153 (1973).

82. R. Olson, *Electron. Packag. Prod.*, 213 (May 1980).

83. M. A. Spivack, J. N. Pike, and W. M. Jayne, *Proceedings of the Electro-Optical Systems Design Conference*, New York, N.Y., Sept. 18–20, 1973, pp. 362–371.

84. N. Pailer and E. Grun, *Planet. Space Sci.* **28**(3), 321 (1980).

85. P. L. Tassano, *J. Vac. Sci. Technol.* **A3**(5), 2036 (1980).

86. P. H. Sheather, *J. Physics E.* **6**(4), 319 (1973).

87. S. L. Kramer and D. R. Moffett, *IEEE Trans. Nucl. Sci.* **NS-28**(3), 2174 (1981).

88. **U.S. Pat. 3,940,530** (Feb. 24, 1976), W. E. Loeb and M. A. Spivack (to Union Carbide Corp.); **U.S. Pat. 3,864,202** (Feb. 4, 1975).

89. U.S. Pat. 4,321,299 (Mar. 23, 1982), R. A. Frosch and R. A. Frazer (to NASA).

W. F. Beach
Alpha Metals